UNIVERSITY OF SOUTHAMPTON

FACULTY OF MEDICINE, HEALTH AND LIFE SCIENCES
SCHOOL OF BIOLOGICAL SCIENCES

THE X-RAY STRUCTURE OF THE SOUTHAMPTON NOROVIRUS 3C PROTEASE LINKED TO AN ACTIVE SITE-DIRECTED INHIBITOR AT 1.7Å RESOLUTION

by

ROBERT JOHN HUSSEY

A Thesis submitted for the Degree of

DOCTOR OF PHILOSOPHY

September 2007

ABSTRACT

FACULTY OF MEDICINE, HEALTH AND LIFE SCIENCES

SCHOOL OF BIOLOGICAL SCIENCES

Doctor of Philosophy

THE X-RAY STRUCTURE OF THE SOUTHAMPTON NOROVIRUS 3C PROTEASE LINKED TO AN ACTIVE SITE-DIRECTED INHIBITOR AT 1.7Å RESOLUTION

by Robert John Hussey

*Noro*viruses are the most common cause of non-bacterial gastroenteritis in humans affecting several million persons world-wide per annum. Essential in the life cycle of any *Noro*virus is a 3C-protease responsible for the processing of a polyprotein encoded by ORF 1 of the viral genome. An understanding of the action, specificity and structure of the 3C-protease is important for a full understanding of the viral life-cycle.

The gene encoding the 3C-protease from Southampton Norovirus (SV) was cloned and expressed using a recombinant strain of *Escherichia coli*. The resulting recombinant protease self-excised from the primary translation product and was purified to homogeneity in milligramme quantities. Substrate specificity of the protease was probed using a colourimetric assay in the form of a series of chromogenic peptides consisting of a yellow chromophore, *p*-nitroaniline, linked to the *C*-terminus of a substrate derived peptide chain of between 3 and 6 residues in length mimicking residues of the natural ORF1 polyprotein substrate P1 to P6 residues.

Crystals of the SV 3C-protease (SV3CP) were obtained and an X-ray structural solution was sought, initially using molecular replacement and subsequently, by incorporating selenomethionine in place of the five methionine residues as a phase reference. A structural solution was finally achieved by MAD after the selenomethionine enzyme had been modified using a synthetic peptide linked to a Michael acceptor inhibitor, Ac-Glu-Phe-Gln-Leu-Gln-propenyl-ethyl-ester. This inhibitor specifically and irreversibly modified the active site cysteine 139 with the amino acids Glu, Phe, Gln, Leu and Gln occupying the S5, S4, S3, S2 and S1 binding sites, respectively. Two other active site amino acid residues, histidine 30 and glutamate 54 are positioned near cysteine 139 indicating the presence of a triad of catalytic residues. The X-ray structure at 1.75 Å resolution allows the interactions of the inhibitor's peptide portion to be defined at the SV3CP active site. Such information may be useful in further development of the Michael acceptor inhibitor or in the design of future prophylactic therapeutically viable agents to tackle *Noro*viral infection.

Contents

List of Figures

Chapter 1 – Introduction

Chapter 2 – Methods in X-Ray Crystallography

Chapter 3 - Synthesis of chromogenic peptide substrates and a Michael acceptor peptidyl inhibitor of SV3CP

Chapter 4 – Results

Chapter 5 - Discussion

List of Tables

Chapter 4 – Results

Chapter 5 - Discussion

Abbreviations

Å	Ångström (10^{-10} m)
Ac	Acetyl
Boc	Butoxycarbonyl
CbV	Camberwell virus
CCD	Charged coupled device
CCP4	Collaborative Computing Project Number 4
ChV	Chiba virus
CNS	Crystallography and NMR System
DCM	*Di*chloromethane
DIBAL	*diiso*butyl aluminium hydride
DIC	N,N-diisopropylcarbodiimide
DIPEA	*di*-isopropyl ethylamine
DM	Density modification
DMF	N,N-dimethylformamide
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
E. coli	*Eschericha coli*
ESI-oa-TOF-MS	Electro-spray ionisation orthogonal acceleration time of flight mass spectrometry
ESRF	European Synchrotron Radiation Facility
FFT	Fast Fourier transform
Fmoc	9-fluorenmethoxycarbonyl
FOM	Figure of merit
FPLC	Fast protein liquid chromatography
FUT	Fucosyltransferase
GTA	Glycosyltransferase A
GTB	Glycosyltransferase B
HAV	Hepatitis-A virus
HBGA	*histo*-blood group antigen
HCV	Hepatitis-C virus
HEPES	4-(2-hydroxethyl)-1-piperazineethanesulphonic acid
HF	Hydrogen fluoride
HIV	Human immunodeficiency virus
HIV-pro	HIV protease

HOBt	N-hydroxybenzotriazole
HPLC	High performance liquid chromatography
IPTG	*iso*-propyl-β-galactopyranoside
LB	Luria broth
Le	Lewis antigen
LV	Lordsdale virus
MA	Michael Acceptor
MAD	Multi-wavelength anomalous dispersion
MALDI-Q-TOF-MS	Matrix assisted laser desorption ionisation quadrapole time of flight mass spectroscopy
MAPI	Michael acceptor peptidyl inhibitor (Ac-EFQLQ-propenyl-ethyl-ester)
MBHA	4-methylbenzhydrylamine
MIR	Multiple isomorphous replacement
MIRAS	Multiple isomorphous replacement anomalous scattering
MR	Molecular replacement
MPD	2-methyl-2,4-pentanediol
NaHMDS	sodium bis(trimethylsilyl)amide
NMR	Nuclear magnetic resonance (spectroscopy)
NoV	Norwalk virus
NV's	*Noro*-viruses
OD_{600}	Optical density
ORF	Open reading frame
PCR	Polymerase chain reaction
PEG	Polyethyleneglycol
pH	$-\log_{10}$ hydrogen ion concentration
pI	Isoelectirc point
*p*NA	*para*-nitroaniline
RBF	Round bottomed flask
RE	Rotary evaporator
RNA	Ribonucleic acid
RTI	Reverse transcriptase inhibitor
RTP	Room temperature and pressure
PI	Protease inhibitor
pSV3C	Expression plasmid construct used for recombinant expression of SV3CP

SAD	Single-wavelength anomalous dispersion
SDS PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
SIR	Single isomorphous replacement
SIRAS	Single isomorphous replacement anomalous scattering
SPP	Signal peptide protease
SPPS	Solid phase peptide synthesis
SRSV's	Small round structured viruses
SV	Southampton virus
SV3CP	Southampton virus 3C-like protease
TEA	Triethylamine
TBTU	O-(Benzotriazol-1-yl)-N,N,N,N-tetramethyluronium tetrafuoroborate
TFA	Trifluoroacetic acid
THF	Tetrahydrofuran
TLC	Thin layer chromatography
TRIS	Tris(hydroxymethyl)aminomethane
UV	Ultraviolet
VLP	Virus-like particle

1.0 Introduction

1.1 The Southampton Virus

Noro-viruses (NV's) previously termed *Norwalk*-like viruses, also termed *small round structured* viruses (SRSV's), have been identified as a major global cause of acute viral gastroenteritis (56). NV's are known to be transmitted via the faecal-oral route, via contaminated food and water, by direct contact, and possibly by transmission through air borne viral particles. Sporadic outbreaks affect all age groups and may occur within a single family, although more commonly occur within confined communities such as hospitals, nursing homes and schools (58). Symptoms are usually present between 1 and 48 hours following infection. Usually the sufferer will experience symptoms of severe gastroenteritis: vomiting, nausea, diarrhoea and abdominal cramps usually lasting up to 72 hours. The loss of fluids and electrolytes through vomiting and diarrhoea are of particular concern in the very young and elderly (59). NV's are fatal only in vulnerable individuals, though in developing countries gastroenteritis is a common cause of death in the under five's where NV's account for just over ten percent of childhood gastroenteritis (68).

The original Norwalk virus (NoV) was first described following an outbreak of gastroenteritis in Norwalk, Ohio, USA during 1968 (4). Bacterial free faecal filtrates derived from infected individuals were orally administered to human volunteers. The volunteers subsequently displayed the same symptoms as individuals infected in the original outbreak, for the first time demonstrating that gastroenteritis could be caused by a non-bacterial agent i.e. a viral pathogen. Since the identification of the Norwalk strain, many outbreaks of non-bacterial gastroenteritis have been attributed to viruses sharing similar morphology to viral particles isolated from the 1968 outbreak (59). By negative stain electron microscopy such viruses are seen to possess an amorphous appearance with a feathery outer edge.

NV's, have attracted considerable research interest with the entire genome sequences for 5 NV's having been described; Norwalk virus (61), Lordsdale virus (LV) (62) Camberwell virus (CbV) (69, 70), Chiba virus (ChV) (71) and the Southampton virus (SV) (63). Comparisons of genome sequence and organisation have placed these viruses in the family *caliciviridae* (64): a viral family consisting of four genera, one of which is the NV's. Other members of the *caliciviral* family include viruses as diverse as: the primate Pan 1 virus (72), Jena bovine virus (73), feline calicivirus (74) and rabbit haemorrhagic disease (75). Interestingly unlike human *caliciviruses*, that cause only gastroenteritis, animal *caliciviruses* are responsible for causing a broad range of conditions. Study of the molecular mechanisms that regulate *caliciviral* replication is of importance to both human and veterinary science.

1.2 Viral structure

The NoV was first visualised following the incipient outbreak of 1968 by immuno-electron microscopy (59). By the early 1990's advances in molecular cloning allowed for the sequencing and expression of the capsid protein, with the first 3-dimensional capsid reconstruction achieved by 1994 using a combination of electron-cryomicroscopy and computer based image processing techniques (65). The viral capsid was shown to consist of a single protein of 58 kDa. The capsid protein arranges into dimers to form a total of 90 capsomeres resulting in a viral particle of 405 Å in diameter possessing 32 large surface indentations 90 Å wide and 50 Å deep. The closely related LV and SV show the same capsid structure (65).

NV's lack a capsid envelope, and this may explain why they are able to survive the strongly acidic conditions of the stomach. Capsid envelopes are generally degraded in the stomach's acidic conditions; therefore those viruses that possess and rely on a capsid envelope are liable to become unstable and therefore unviable on entering the stomach. Since NV's have evolved to not require a capsid envelope they remain relatively stable in the stomach allowing them to remain viable and capable to infect (76).

1.3 Noroviral infection

SV infects and replicates in enterocytes (17); a type of epithelial cell lining the small and large intestine. This leads to infective viral particles being shed in the faeces; hence poor hygiene is responsible for spread of the virus *via* the oral-faecal route. Research focusing on susceptibility and resistance to infection has been confined to the prototype Norwalk strain but should serve as an adequate model for all NV's.

1.3.1 Susceptibility and Resistance to NV infection

Studies have shown short term immunity is provided after initial infection (23,55). In 1974 Wyatt (55) demonstrated that individuals infected by NV displayed a resistance 6 – 14 weeks post NV exposure. However, conflicting studies have reported that the possession of antibodies against NV is not correlated with protection against further NV challenge (2), and in some cases, individuals displaying high levels of antibody actually showed an increased susceptibility to infection compared to those with either no antibody or low levels of antibody (23).

To assess long term resistance to re-infection with NV, Parrino *et al* (23) exposed 12 individuals to NV and of these 12 only 6 individuals displayed clinical symptoms of infection. All 12 were re-exposed 27-42 months later and the 6 individuals that had displayed *no* symptoms following the initial exposure again displayed *no* symptoms following the second NV challenge, suggesting that some individuals possess an innate resistance to NV infection. Of the 6 individuals that *had* displayed symptoms of infection following the initial challenge, all again developed symptoms following the second NV challenge, suggesting they possessed no long term immunity to NV infection.

So it appears that some individuals possess an innate resistance to NV infection (21, 23), interestingly such individuals appear in familial clusters (22). Of those that *do* suffer NV infection, most experience some short term resistance, though none enjoy long term resistance to re-infection. So what are the underlying mechanisms governing an individuals disposition to NV infection? The NV is known to bind to enterocyte cell surface antigens. These antigens have been shown to be *histo*-blood group antigens (HBGAs) (13, 12). HBGAs are genetically determined and so vary between individuals, correspondingly this variation may define who *is* and who *is not* susceptible to NV infection.

1.3.2 The ABO, Lewis, and 'secretor type' blood group systems

Genetically determined blood groups: A, B and O, have been linked to susceptibility of infection of several viruses (8, 16), including the *Norwalk* virus (8, 12). The ABO blood group was first described by Karl Landsteiner in 1901 following the study of human red blood cells. Landsteiner observed two types of red blood cell surface antigens ("blood group antigens") naming them A and B giving rise to four blood groups: i) type A: those with the A-antigen; ii) type B: those with the B-antigen; iii) type AB: those with both A- and B-antigens, and; iv) type O: those possessing an unmodified precursor antigen of the A- and B-antigens. It was later discovered that blood group antigens are not confined to just the surface of red blood cells but are also secreted by most mucus secreting mucosal cells such as those of the gut, and as such are present in biological fluids including saliva (4). Further research has linked the blood group antigens A and B with other antigens such as the *Lewis antigens*. These observations led to a change in name from 'blood group antigens' to '*histo*-blood group antigens' (HBGA's) (15).

HBGA's are complex carbohydrates linked to cell surface proteins or lipids via N- or O-linked glycans. HBGA's are involved in recognition of self and non-self, cell to cell interactions, and of relevance to this thesis, play a role in the binding of some viral particles to host cell surfaces that lead to viral internalisation and infection (7). In this way an individual's blood type can, for some viruses, affect their susceptibility to infection.

HBGA's result from the modification of a precursor antigen common to all blood types: the "*H-type precursor antigen*". An individual's blood group is dependent upon expression of a number of enzymes that each play a role in a cascade of possible modifications to the *H-type precursor antigen*. Inactivating mutations can occur in a number of these enzymes and it is therefore the level of modification of the initial *H-type precursor antigen* that defines an individual's blood type. The lowest level of modification involves the fucosylation of the *H-type precursor antigen* by the enzyme: α1-2-fucosyltransferase. α1-2-fucosyltransferase is encoded by the *FUT2* gene (18) that mediates the addition of a fucose sugar to the *H-type precursor antigen* to yield the mature *H-antigen*. In some individuals the *FUT2* gene contains an inactivating mutation preventing the fucosylation of the *H-type precursor antigen* (19). Without fucosylation of the *H-type precursor antigen* this relatively immature HGBA is not secreted by mucosal cells, so individuals possessing the inactivating *FUT2* mutation are referred to as *non-secretors*. In contrast, those that do possess a functioning *FUT2* gene will permit fucosylation of the *H-type precursor antigen* resulting in the mature *H-antigen* being secreted by mucosal cells, such individuals are referred to as *secretors*. If no further modification of the *H-antigen* occurs, an individual will possess an "O" type blood group (15).

Two enzymes are encoded by the ABO locus, these are: glycosyltransferase A (GTA) and glycosyltransferase B (GTB). GTA mediates the addition of a *N*-acetylgalactosamine to the *H-antigen* giving rise to the *A-antigen* resulting in blood group type A; GTB mediates the addition of a galactose sugar to the *H-antigen* giving rise to the *B-antigen* resulting in blood group type B (1). Individuals expressing a functional version of both GTA and GTB will possess both the *A*- and *B-antigens* and a blood group type AB.

In addition to ABO blood type and secretor status, additional fucosylations of the *H-precursor antigen*, *H-antigen, A-antigen* and *B-antigen* gives rise to the *Lewis* blood type. In secretor negative individuals, a fucose group is added to the *N*-acetylglucoseamine of the *H-type precursor antigen* in either a α1,4 or α1,3 linkage giving rise to the Lewis antigens: Le^a or Le^x respectively. Similar additions via α1,4 or α1,3 linkages to the *H-antigen* of secretor positive individuals gives rise to the Le^b or Le^y antigens respectively. Fucosylations of the *A-antigen* yield the Lewis antigens ALe^b or ALe^y. Similar fucosylations of the *B-antigen* yield the Lewis antigens: BLe^b or BLe^y. Several genes have been identified that encode the fucosyltransferases responsible for the fucosylations giving rise to the various Lewis groups, the most implicated are the: *FUT3* and *FUT5* genes (15). As with inactivating mutations of the *FUT2* gene defining secretor status, inactivating mutations of the *FUT3* and *FUT5* genes define the Lewis blood type (15).

1.3.3 Secretor status and susceptibility to NV infection

Several studies have reported a correlation between secretor status and susceptibility to NV infection. A secretor negative phenotype appears to favour resistance to infection. In volunteer studies where individuals were challenged with NV, secretor negative individuals were always found to be resistant to infection (2, 9). Further, virus-like-particles (VLP's) have been shown to specifically bind *to H-type* HBGAs that are present only in secretor positive individuals (11, 8, 12, 3, 9, 2). Further still, non-secretors have been shown to have lower NV antibody titers than secretors, suggesting these individuals are less prone to infection by NV (1).

Secretor status, as discussed, is determined by the *FUT2* gene. The *FUT2* gene is encoded for by a single 999bp exon. Several mutations in *FUT2* have been identified that lead to this gene's inactivation. A G428A mutation is found in Caucasian populations at a frequency of approximately 30%, and an A385T in Asian populations with a frequency of approximately 20% (14). It has been demonstrated that NV must to bind to the *H-antigen HGBA* before internalisation into the host enterocyte (12), therefore only secretor positive individuals can suffer NV infection since it is only these individuals that secrete the *H-antigen* from the epithelial cells of the gut. In those individuals that are secretor positive but were still not infected, it is thought that pre-existing antibodies resulting from recent NV infection must have played a role in resistance (9).

1.3.4 ABO blood group and susceptibility to NV infection

In addition to secretor status, ABO blood group has been linked to an individual's risk of infection by NV. Studies have demonstrated that those expressing the *A-* or *B- antigen* are less likely to be infected than those expressing neither and therefore of blood type O (2, 13, 9, 6). This is presumably due to the more extensive modification of the *A-* and *B-antigens* over the *H-antigen*, resulting in a reduced ability to bind NV. To further delineate the effect of blood group on risk of infection, VLPs have been shown to bind more strongly to *A-* and *H-antigens* (H-antigens are possessed by blood type O individuals) than to the *B-antigen* (11, 8, 9, 3). Finally, the presence of the *B-antigen* has also been linked with lack of immune response to NV (6) and may also be due to the more extensive modifications of the *A-* and *B-antigens* reducing NV binding (13, 9, 6).

In summary, although results clearly show that individuals of blood group O are more susceptible to infection, expression of the *B-antigen* does not entirely prevent binding; hence secretor status is the best predictor of susceptibility to the NV (2, 9).

1.3.5 Lewis group and susceptibility to NV infection

There is also some evidence that the Lewis blood group may affect susceptibility to NV infection. There is evidence that NV particles bind to Le^b antigens but not Le^a (11). In addition, Le^a individuals have been shown to have a much lower antibody titre to NV than Le^b individuals. In contrast, other studies have reported no effect of Lewis type on binding (3) or antibody titre when comparing Lewis positive and Lewis negative individuals (1). The disagreement in results is thought to be reflective of the individual's secretor status rather than Lewis type (1).

1.3.6 NV binding to HBGA's is strain specific

The binding of NV to HBGA's appears to be strain specific. For example, infection with the *Snow-Mountain Noro*virus has been shown to be independent of secretor status (5) and instead depends upon the presence of the *B-antigen* (11). Further research is required to fully define binding of NV particles to host antigens and therefore determine an individual's susceptibility to infection.

1.4 SV genome structure

The absence of an *in vitro* cell expression system for NV's, and the relative fragility of the viral genome has presented problems for molecular cloning approaches and generally hindered progress in study of NV's. Early work by Jiang *et al* (61) resulted in the generation of a cDNA library that allowed for the identification of a viral RNA-dependent RNA polymerase by the comparison of known sequence motifs. Work by Matsui (77) focused on the expression of a cDNA fragment obtained by amplification of a cloned region of the NV genome in *E. coli.* This cDNA was transformed into *E. coli* using the *bacteriophage λ vector* and the expressed fragment was shown to react with sera raised from individuals infected with NV.

Use of direct reverse transcription PCR sequencing from the faeces of infected individuals resulted in the elucidation of the entire genome sequence for the NV (79) and two UK born strains SV (62) and LV (80).

The NV genome consists of a molecule of single stranded positive sense RNA of 7.4-7.7kb in size with a polyadenylated 3' terminus. The genome comprises three open reading frames ORF 1, 2 and 3. ORF 1 is positioned at the 5' terminus and codes a large non-structural polyprotein. ORF 2 follows ORF1 and codes the single capsid protein. ORF 3 follows ORF2 and codes for a small basic protein that may assist in the assembly of newly synthesised viruses. *Caliciviruses* display a unique genome organisation in comparison to other positive sense RNA viruses such as the poliovirus, by not coding the capsid protein at the 5' region of the genome (64).

Phylogenic analysis of *Caliciviral* genomes indicates the presence of four subgroups of *Caliciviruses*. Further genome sequence comparisons of viruses from all four groups reveals the presence of two distinct genogroups (64); genogroup I and genogroup II. A difference in genome size exists between the two groups; the genome of genogroup II being approximately 200 nucleotides longer than genogroup I. The size and overlap of the three ORF's also differs between groups I and II. Based on genome sequence analysis NoV and SV can be placed in genogroup I, and ChV, CbV and LV in genogroup II. To demonstrate genome differences, comparison of the 5' region of LV and the equivalent region in the NoV and SV reveals considerable sequence variation. Further, ORF1 is considerably smaller in LV than NoV and SV; this coupled with the ORF1 sequence diversity suggests some difference in regulatory signals and genome secondary structure between the two groups. However, these differences are relatively subtle resulting in the genome structure of each genogroup still remaining quite similar. Downstream of the 5' region of ORF1 both genogroups display sequence motifs characteristic of the *Picornaviral* 2C-helicase, 3C-protease and 3D-RNA-dependent RNA-polymerase (64). In both genotypes ORF2 is frame shifted relative to ORF1 and the 5' terminus of the capsid protein coded by ORF2 overlaps the 3' terminus of the 3D-RNA-dependent RNA polymerase. The size of the overlap differs between the genogroups. Genogroup I viruses show a 17 nucleotide overlap; genogroup II viruses a 20 nucleotide overlap. Further reading frame overlap exists in both genogroups, where the initiation codon of ORF 3 overlaps ORF2 by just one nucleotide, resulting in ORF3 being in a different reading frame to ORF 2 but the same as ORF1 (64) (*Figure 1.1*).

1.5 Proteolytic processing of the SV ORF 1 polyprotein product by a virally encoded protease

Early research efforts using a cDNA clone of the entire SV genome permitted the translation of ORF1 in a rabbit reticulocyte lysate expression system. Translation of ORF1, occurring from one of three tandem initiator codons, resulted in the production of three major non-structural protein products (56). The three products were: a 48 kDa *N*-terminal protein (P48), a 41 kDa 2C-like helicase, and a 113 kDa *C*-terminal protein. Since none of the three proteins matched the predicted mass of approximately 200 kDa of the ORF1 gene product, a proteolytic cleavage event was assumed to have taken place, most probably co-translationally, where a precursor polyprotein i.e. the 200 kDa protein corresponding to the translation of the entire length of ORF1 was cleaved by a host cell or virally encoded protease into the three proteins observed. Analysis by site directed mutagenesis identified a region of ORF1 that coded for a 3C-like protease, this confirmed that the proteolytic cleavage of the 200 kDa polyprotein was completed by a virally encoded protease not a host cell protease. The 3C-like viral protease was suggested to belong to a group of proteases known as the *chymotrypsin-like cysteine proteases*, a subset of the *cysteine protease* family characterised by an active site catalytic cysteine residue. Site directed mutagenesis revealed the 3C-like protease to be the only protease coded by the viral genome as mutations of the region of ORF1 containing the 3C-like protease motif resulted in the production of a single 200 kDa protein indicating that the ORF1 protein product had undergone no proteolytic processing. Further mutagenic studies located the exact cleavage sites recognised by the SV 3C-like protease (SV3CP): cleavage was identified between a Q/G di-peptide at the both the *N*- and *C*- terminus of the 41 kDa 2C-like helicase region (56) (*Figure 1.1*).

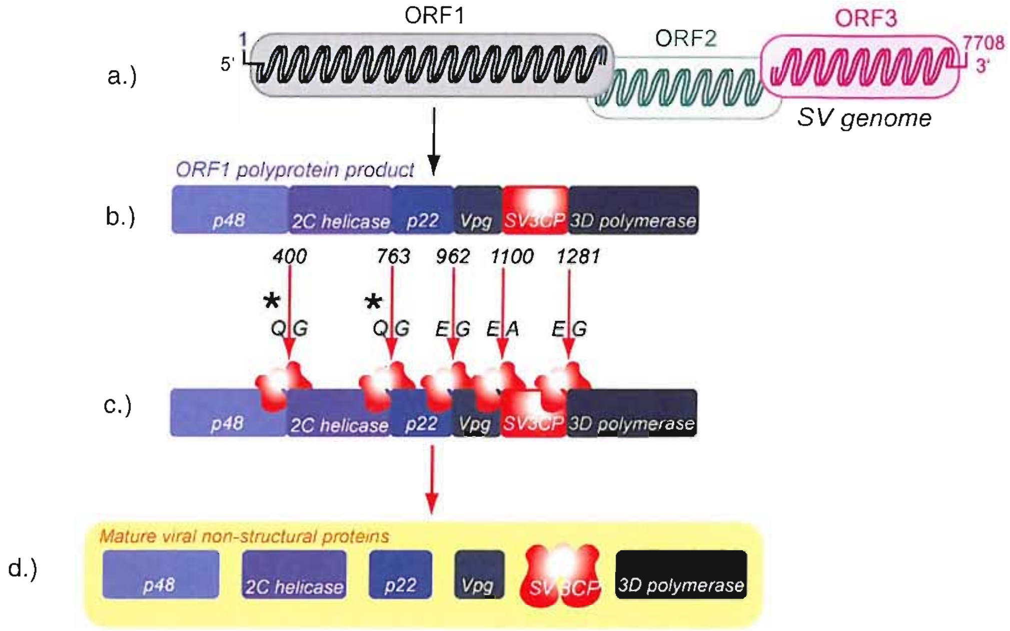


Figure 1.1: Proteolytic processing of the SV ORF 1 polyprotein product by SV3CP. *a.) The genomic arrangement of the three open reading frames of the Southampton virus genome; ORF1, ORF2, and ORF 3. b.) The ORF1 polyprotein product. c.) The points of cleavage of the ORF1 polyprotein by SV3CP. Numbering indicates the residue C-terminal to the cleaved peptide bond. *Indicates the initial cleavage events to yield the 41 kDa and 113 kDa proteins. d.) The mature protein products following complete cleavage of the ORF1 polyprotein by SV3CP.*

In vitro expression of the ORF1 polyprotein in *E. coli* revealed three further sites of cleavage within the 113 kDa protein fragment (58) giving rise to an additional four proteins: a 22kDa protein similar to the *Picornaviral* 3A-protein (P22), a 16 kDa viral genome linked protein: VPg, a 19 kDa protein: SV3CP itself, and a 57 kDa RNA-dependent RNA-polymerase: 3D Pol. Sequencing of the *N*- and *C*-terminals of each of these four proteins allowed the cleavage sites of SV3CP to be fully defined. SV3CP shows a preference for cleaving at: LQ/GP and LQ/GK to generate the three initial products of the 200 kDa precursor polyprotein, then at: ME/GK, FE/AP and LE/GG in the 113 kDa polyprotein (*Figure 1.1*). It has been proposed that SV3CP will preferentially accommodate a glutamine residue at the active site S1 position, as it does during the cleavage of the 113 kDa polyprotein, but will also accept a glutamate residue when cleaving the 200 kDa polyprotein product of ORF1. A difference in rate of substrate cleavage by SV3CP may occur depending upon the substrate residue occupying the SV3CP S1 site; it has been suggested that this may be part of a mechanism to mediate rates of cleavage as a control of the production of protein precursors i.e. the 113 kDa polyprotein. This could be of importance during viral assembly since regulation of the rate of production of precursor proteins that possess modified enzyme activity may control rate of translation and viral replication. It has been demonstrated that SV3CP is closely involved with the 57 kDa 3D-Pol (57); possibly to control the rate of each others activity, however this has not yet been satisfactorily demonstrated.

Since the proteolytic processing of the initial 200 kDa precursor polyprotein and subsequent cleavage of the 113 kDa polyprotein are both essential to yield functional viral proteins, SV3CP presents itself as a particularly viable target for antiviral strategies. Halting the activity of SV3CP will arrest viral replication.

1.6 Viral proteases as a therapeutic target of antiviral compounds

Therapeutic agents capable of arresting viral replication were first identified over 50 years ago (81). These early drugs were largely nucleoside analogues, designed to inhibit viral DNA synthesis, and in the case of retroviruses: reverse transcription of the viral genome. With the notable exception of the guanine analogue: *acyclovir* (82) that is still used extensively as an antiviral (mainly in the treatment of the herpes simplex virus) these early nucleoside analogues often displayed limited clinical efficacy, and caused adverse side effects. Combined with the appearance of drug resistant viral strains, most notably those of HIV (83), there was pressure for the development of a new generation of anti-viral compounds that acted upon more virus-specific targets.

By the early 1990's a deeper understanding of the viral life cycle identified a number of key viral enzymes that played vital roles in viral replication. It was proposed that if the activity of these key enzymes could be inhibited there would be a corresponding cessation in viral replication. Chief among these newly identified viral targets were virally encoded viral proteases (28), responsible for the proteolytic processing of non-structural viral polyproteins to yield the functional, discrete viral protein products, and also, the processing of structural viral polyproteins required during the assembly of newly synthesised viral particles. Virally encoded proteases presented such an attractive potential target they quickly became the focus for a new generation of virus-specific antiviral agents. Developments in design and synthesis of drugs targeted against HIV, Hepatitis-C virus and the Human Rhinovirus will now be discussed as they represent areas of intense research within this field.

1.6.1 The role of protease inhibitors in the treatment of HIV infection

The first anti-retroviral agents approved for the clinical treatment of HIV infection were nucleoside analogues designed to target *HIV reverse transcriptase,* however as with similar nucleoside analogues developed during the 1950's, these drugs, as part of a mono-therapy regime showed only moderate efficacy, and displayed adverse side effects (29). A second generation of *reverse transcriptase inhibitors* (*RTI's*) were structurally unrelated to the nucleoside analogues that preceded them and as such were termed *non-nucleoside reverse transcriptase inhibitors*. *Non-nucleoside RTI's* were shown to be effective in slowing HIV replication *in vivo* initially, though unfortunately their efficacy in the clinical treatment of HIV infected individuals was subsequently tempered by the development of drug resistant HIV strains (30). It is the propensity of HIV to develop these drug resistant strains that required a less mutationally prone viral target to be sought. Investigation showed that mutations within the *HIV protease* (*HIV-pro*) resulted in non-infectious viral particles (54), and as such *HIV-pro* presented itself as an ideal target for an antiviral agent. *HIV-pro* belongs to the aspartic protease family and is responsible for the proteolytic processing of the *gag* polyprotein to yield 4 *gag* structural proteins, and also the proteolytic processing of the *pol* polyprotein to yield 3 non-structural proteins: the *HIV reverse transcriptase*, an integrase and the *HIV-pro* itself (30) (as with many viral protease a poorly understood autocatalytic event must occur that allows for the self-excision of the protease from its precursor polyprotein).

There are presently nine approved protease inhibitors (PI's) that have been used, or are used clinically in the treatment of HIV infection. These include the first generation PI's: *Saquinavir, Ritonavir, Indinavir, Nelfinavir* and *Amprenavir*, and the second generation PI's: *Fosamprenavir, Lopinavir, Atazanavir* and *Tipranavir* (26). First generation PI's were peptide based and affected their inhibitory action by including a non-hydrolysable amide bond between residues that occupied the S1 and S1' substrate binding sites of *HIV-pro* (33). Though these PI's arrested viral replication *in vitro* (34) due to their highly peptide nature they displayed a low bio-availability. Despite this, in clinical trials, they were showed to slow viral replication as reflected in a reduced *HIV*-RNA level in the plasma and increased CD4 counts (35) of the infected patient. Development of these first generation PI's made them less peptide like to increase bioavailability. HIV PI's are not without their own side effects, amongst them are: diarrhoea, dyslipidemia, and increased risks of cardiovascular disease and diabetes (84). These second generation PI's in combination with RTI's currently provide the most effective approach in the suppression of HIV replication (85).

1.6.2 The role of protease inhibitors in the treatment of Hepatitis-C infection

Infection by Hepatitis-C (HCV) causes inflammation of the liver: *hepatitis*. Although in the short term infection is often asymptomatic, in the long term sufferers may develop chronic hepatitis leading to liver cirrhosis (scarring of the liver) and an associated increase in the risk of liver *carcinoma* (87). Currently HCV is treated through a combination of *peginterferon-α 2a* and the nucleoside analogue *ribavirin* (86). This treatment strategy shows a 40% success rate in curing individuals infected with HCV-1 serotype, and a 80% success rate with those infected with the HCV-2 serotype (44). However, since treatment with peginterferon-*α 2a* and *ribivarin* is not always successful and even when it is, may produce adverse side effects (88), the need for an alternative antiviral agent is evident, and the two virally encoded HCV proteases present an ideal target.

HCV is a small, enveloped RNA virus belonging to the family of viruses: *Flaviviridae* (43). The HCV genome is comprised of a single strand of positive sense RNA of approximately 9.6 kB in length that encodes a single polyprotein precursor. The cleavage of this polyprotein precursor differs from the equivalent cleavages in the life cycles of HIV, HRV or SV, in that 5 of the ten proteolytic events to yield mature viral proteins are mediated by host cell *signal peptide proteases* (SPP's). The remaining 5 proteolytic events are mediated by the actions of the NS2/3 protease (45, 46) or the NS3 protease alone (47). The major focus has been placed on development of an inhibitory compound of NS3. The X-ray crystal structure of NS3 (48) allowed it to be placed in the *trypsin-like serine protease* family, and has shown that NS3 possesses a shallow active sight binding cleft in comparison to other viral serine protease and cellular serine proteases. This presents an obstacle to rational knowledge led design of NS3 inhibitors since the number of substrate/inhibitor to NS3 active site cleft interactions is relatively small. However due to a number of differing stereo-chemical requirements between NS3 and host cell serine proteases, NS3 still present itself as a viable target of an antiviral agent (49). Despite challenges associated with the identification of such a compound, an inhibitor of HCV NS3: VX-950, is shortly to be introduced (scheduled for release 2009) for the clinical treatment of HCV infection (50). VX-950 is a modified substrate derived peptide inhibitor of NS3 with an α-ketoamide as the functional moiety that is capable of forming a covalent but reversible bond to the NS3 active site serine and thereby inhibit NS3 protease activity (50). VX-950 has proved successful in treatment of HCV infection during clinical trials reflected in a significant decrease in HCV-RNA levels in the plasma of infected patients (51).

A second inhibitor of NS3, BILN-2061 (trade name: *Ciluprevir*) is also anticipated to be released for clinical use shortly. BILN-2061 is a substrate derived hexa-peptide (amino acid sequence: DDIVPC) competitive non-covalent inhibitor of NS3 that has displayed good affinity and high potency in *in vitro* studies (52) and proved well tolerated by infected patients in clinical studies causing a rapid decline in viral load (53). So despite being largely peptide in nature, a characteristic usually associated with poor bioavailability, BILN-2061 appears an efficacious anti-HCV agent.

1.6.3 The role of protease inhibitors in the treatment of Human Rhinovirus infection

A member of the diverse *Picornavirus* family of viruses, the human rhinovirus group of viruses comprised of more than 100 serotypes, has been identified as the major pathogenic agent responsible for mild upper respiratory tract infections (the common cold) (38). Aside from compounds that target viral attachment to host cell surface receptors, and others that prevent subsequent viral un-coating (36), attention has been focussed upon the virally encoded 2A (HRV-2A) and 3C proteases (HRV-3C) (37). The human rhinoviral genome encodes a single polyprotein that is initially cleaved by the HRV-2A protease to yield a precursor capsid protein, and two further polyproteins that are comprised of non-structural viral proteins. Further proteolytic processing of the precursor capsid protein and the non structural polyproteins is mediated by the HRV-3C protease that either acts independently or in complex with a viral RNA-dependent RNA-polymerase: 3CD. It is the activity of the HRV-3C protease that has been the focus of attention in design of an anti-HRV agent.

1.6.3.1 The HRV-3C protease and Michael-acceptor substrate derived peptide inhibitors

The HRV-3C protease is a cysteine protease that displays a trypsin-like serine protease fold (39) (together these properties distinguish it from other viral proteases) that cleaves its polyprotein substrate between the di-peptide: Gln/Gly. It has been shown that the preferred substrate amino acid sequence for cleavage by HRV-3C is: Leu, Phe, Gln that occupy the HRV-3C active site at the S3, S2 and S1 positions respectively (41). Modified peptide inhibitors that include the preferred amino acid recognition sequence but possess a *C*-terminal chemical moiety capable of reacting with the HRV-3C protease active site cysteine residue, have been developed that in *in vitro* studies completely inhibit HRV-3C protease activity (40, 42). One such modified peptide inhibitor includes a Michael-acceptor type group at its *C*-terminus, which undergoes nucleophilic attack by the HRV-3C protease active site cysteine thiol group resulting in the inhibitor becoming covalently and irreversibly bound to the active site cysteine and therefore abolishing protease activity (40). Development of these peptide-Michael-acceptor inhibitors has led to compounds that not only inhibit HRV-3C protease activity *in vitro* but also display antiviral properties *in vivo* (89). The properties of these Michael-acceptor inhibitors along with their mode of action and method of synthesis is discussed in more detail in *Sections 3.5 – 3.11* of this thesis, as it was this type of cysteine protease inhibitor that was the basis for the design and synthesis of an inhibitor of SV3CP.

1.7 Research aims

At the outset of this research project there were four clear aims:

1.) The recombinant expression, purification to homogeneity, and crystallisation of SV3CP.

2.) The synthesis of a series of substrate derived chromogenic peptides for use in a colorimetric assay of protease activity in order to probe substrate SV3CP specificity and so provide the basis for a peptide based inhibitor of SV3CP.

3.) The design and synthesis of such a peptide based SV3CP inhibitor suitable for development as a therapeutically viable anti-SV agent.

4.) Crystallisation of SV3CP in complex with the peptide based inhibitor to allow dissection of interactions governing substrate/inhibitor binding at the SV3CP active site cleft, and so provide further information on which to develop a therapeutically viable anti-SV agent.

2.0 Methods in X-Ray Crystallography

2.1 X-Ray crystallography: an overview

X-Ray crystallography is the most widely applied and successful technique available to the biochemist seeking an accurate depiction of protein structure. Currently the Protein Data Bank (102, 128) holds co-ordinates describing over 35, 000 protein and nucleic acid structures that have been determined by X-ray crystallography. A crystallographic approach to protein structure solution provides the user with a highly accurate 3-dimensional model describing structural features from broad detail such as global folds to the minute details of atomic interactions. Further investigation of such protein models can often reveal information regarding the target protein's mechanism of activity; how it may bind ligands and other proteins; evolutionary relationships between proteins; and, in the case of enzyrnic proteins, information detailing substrate binding and turnover - often very useful in aiding drug design and development.

X-Ray crystallography is a powerful technique however some major hurdles have to be overcome by the crystallographer to solve a protein structure. The first is a supply of a relatively large amount of highly pure protein, typically >98% purity is essential. Secondly, crystals of the target protein must be grown, an often difficult and time consuming process. Under certain environmental conditions many molecular substances including proteins can crystallise, an orderly three dimensional array of protein molecules held together by non-covalent interactions. To be of use these crystals must be of a high quality and able to scatter X-rays to yield information to a minimum resolution of around 3.5 Å to allow unambiguous positioning of protein atoms. The final major hurdle facing the crystallographer is to assign a phase angle to the X-rays scattered by the protein crystal to allow interpretation of the X-ray data and the construction of a model representative of the target protein (105, 106, 107). Each of these issues will be dealt with in more detail during this chapter.

X-Ray crystallography relies upon the phenomenon of X-ray scattering by protein crystals. X-Rays, as with visible light, are waves of electromagnetic radiation. Electromagnetic radiation belonging to the visible spectrum and the X-ray spectrum differs only in wavelength; visible light possesses a wavelength of between 3x10^{-9} m and 9x10^{-9} m; the X-ray spectrum covers electromagnetic radiation of wavelength between 1x10^{-11} m and 1x10^{-9} m (105, 106, 112). It is this comparably short wavelength that allows X-rays to interact and be diffracted by the electron cloud of an atom similar in size to the X-ray wavelength - typically around 1.0 Å in diameter (1.0 Å is equivalent to 1x10^{-10} m or 0.1 nm). The relatively long wavelength of visible light does not permit this type of scattering, resulting in even the most powerful microscopes only being able to resolve detail separated by 50 Å. Given that the average inter atomic distance is between 1-3 Å it follows that electromagnetic radiation of a wavelength of approximately 1.0 Å must be used to allow imaging of a molecule at the atomic scale - hence the need to use X-rays (105, 106, 107).

X-Ray crystallography is not a direct imaging technique, X-rays cannot be focussed as can visible light by a lens to produce an image. The unique pattern of diffracted X-rays resulting from the interaction of an X-ray beam and a protein crystal must be recorded and later undergo complex computer based calculations in order to reconstruct a map detailing the electron density of the protein arranged in the crystal. This computational approach is often referred to as a 'virtual lens' as the patterns of diffracted X-rays recorded by the crystallographer are analogous to the arrangement of visible light rays at the point of the lens before being focussed to form an image. Atoms can then be modelled into the experimentally determined electron density maps to obtain a reliable model of the molecular structure of the protein (105, 106, 107).

Protein structural determination by X-ray crystallography does have its limitations. Firstly, as relatively little detail is known as to the process of protein crystallisation it is almost impossible to predict the conditions under which the target protein will crystallise, requiring a time consuming trial and error approach. Also, the final structural model explains little of the dynamic behaviour of the protein, since it is a model of the protein held in a crystal lattice and often in conditions paying little similarity to the physiological conditions in which the protein would normally exist. The complementary technique of *Nuclear Magnetic Resonance* (NMR) is useful in providing information revealing the atomic structures of proteins in solution (101), although is rather restricted to relatively small proteins. In the favour of crystallography, highly accurate structural models can be obtained for proteins or complexes of any size and only very rarely have these differed considerably from models obtained of the same protein by NMR.

2.2 Protein crystals

2.2.1 The nature of protein crystals

The need to obtain protein crystals to allow for structural determination is rooted in the weak X-ray diffracting ability of the small and therefore relatively lowly electron dense atoms that comprise protein molecules, mainly; C, O, N, S and H. The ability of a single protein molecule to diffract X-rays is so low as to be undetectable. However, the summation of diffracted X-rays of many protein molecules can be of such a level as to be measurable. To allow for useful diffraction data to be collected the protein molecules have to be held in a rigid and, most importantly, ordered fashion. Protein crystals provide such a system. Unfortunately, unlike a durable salt crystal, where molecules are held together by strong ionic interactions or diamonds where carbon atoms are held in a rigid lattice by covalent bonds, protein molecules are held within a crystal lattice by relatively weak hydrogen bonds, salt bridges and hydrophobic interactions. Coupled with the irregular shape of protein molecules and the associated unsuitability to adopt a closely packed crystal structure, on average 50% of a protein crystal will be made up of solvent (referred to as the 'solvent content'). This results in protein crystals being quite fragile in nature (105, 106, 107).

2.2.2 Arrangement of protein molecules in the crystal lattice

The most basic element comprising a protein crystal is termed the *asymmetric unit*. The asymmetric unit may consist of just a single protein molecule, or of multiple protein molecules. If the asymmetric unit consists of multiple molecules these may be subject to *non-crystallographic* symmetry to form oligomeric complexes; dimers, trimers, tetramers etc (*Figure 2.1*).

Most protein crystals do not consist of simple repetitive translations in three dimensions of the asymmetric unit in a fixed orientation. Protein molecules will usually assemble into more complex arrangements to form a crystal lattice, with the asymmetric unit subject to symmetry operations referred to as *crystallographic symmetry* i.e. the arrangement of multiple asymmetric units by rotations and translations to generate a single *unit cell*. The unit cell is the smallest repeatable unit that may form the crystal lattice *without* undergoing any rotation but only translations in the three dimensions of the lattice *(Figure 2.1)*(105, 106, 107).

The edges of the unit cell are denoted by the letters a, b, c and the internal angles α, β, γ (*Figure 2.1*). Dependent upon these values, protein crystals adopt one of seven crystal systems describing the three dimensional shape of the unit cell; *monoclinic*, *triclinic*, *orthorhombic*, *tetragonal*, *cubic*, *trigonal* and *hexagonal*. On a technical note, the unit *cell* dimensions, a, b, c and α, β, γ can be calculated from a single diffraction pattern. The unique arrangement of reflections (although not their associated intensities) of a diffraction pattern describes the unit cell edge lengths and internal angles, allowing the crystal system to be instantly determined (105, 106, 107).

The arrangement of the asymmetric unit in the unit cell can be quite complex. The corners of the unit cell can be considered to be points of symmetry around which the asymmetric unit may be arranged. If these are the only points of symmetry the crystal is said to possess a primitive lattice. However, it is common to have addition points of symmetry or *lattice points* within the unit cell. In addition to a primitive lattice there are a further three types of centring; C, F and I-centring. Each differs in the number and location of the additional lattice points. In combination with the seven possible crystal systems a total of 28 possible lattice permutations arise, although 14 of these are mathematically redundant and never occur, leaving a total of only 14 possible lattice types. These are known as the 14 *Bravais lattices*. The Bravais lattice describes the crystal system and centring, but explains nothing of the arrangement of protein molecules within the unit cell; this is described by a further level of arrangement - the *point group*. Within the unit cell the asymmetric unit may be symmetrically arranged by a rotation around an axis passing through a lattice point. Theoretically, mirroring of a protein molecule about a point may occur to generate a mirror plane, however, since protein molecules consist solely of the L-enantiomer of amino acids, mirror planes cannot exist within protein crystals. For mirror planes to occur it would be necessary for some protein molecules to consist solely of the D- enantiomer of amino acids as well as some consisting solely of L-enantiomer. Of interest mirror planes do exist within small molecule crystals (proteins are considered macromolecules); fortuitously such an arrangement eliminates the phase problem allowing for structure solution without calculation of phases. Despite this, there exist 32 theoretical symmetrical arrangements of the asymmetric unit resulting from rotations, inversions and roto-inversions. These 32 theoretical arrangements are termed *point groups* (105, 106, 107).

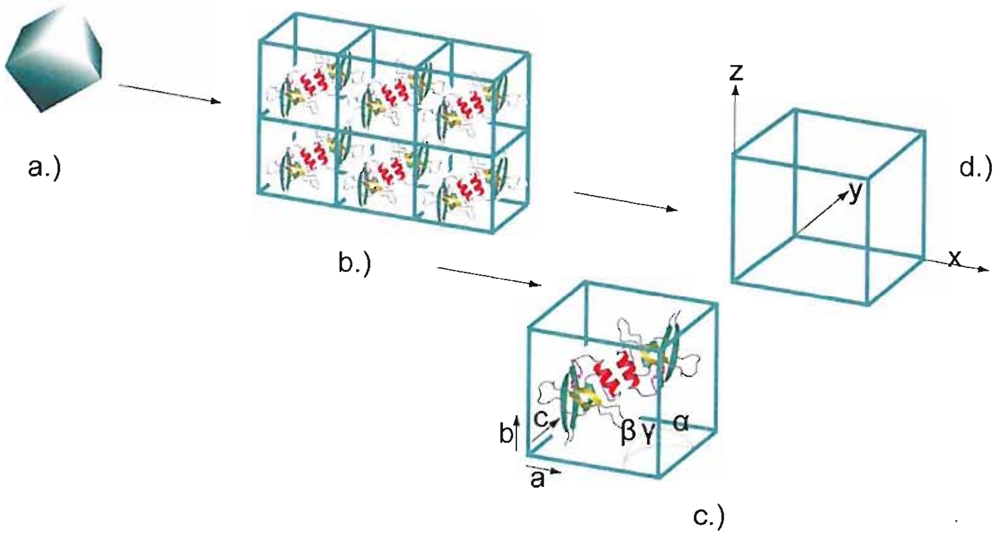


Figure 2.1: Diagrammatic representation of the asymmetric unit, unit cell and crystal lattice. *In this instance the asymmetric unit consists of a single protein molecule shown here with a single alpha helix in red, and two beta sheets, one in green the other in yellow. The unit cell possesses the simplest of crystallographic symmetry with just a 2-fold axis resulting in the asymmetric unit being rotated 180° to it's self. No translational symmetry elements exist in this example. a.) Representation of a protein crystal. b.) Representation of the arrangement of the asymmetric unit and resulting unit cell in the crystal lattice. c.) A single unit cell, showing the simple 2-fold symmetrical arrangement of the asymmetric unit in the unit cell, with unit cell edges a, b, c and internal angles α, β, γ labelled. d.) Representation of the unit cell (protein molecules not shown) with the unit cell edges labelled along which the coordinates x, y, z are measured.*

It is important to note that a point group describes symmetry around a point that remains unchanged i.e. is not subject to any translation. Further symmetry elements are generated by translations of the asymmetric unit in the unit cell. Such an example results from a combination of rotation and translation of the asymmetric unit along an axis resulting in what is termed a *screw axis.* Combination of all types of permitted symmetry elements with the 14 Bravais lattices results in 65 theoretical crystal lattice arrangements; known as *space groups* (105, 106, 107).

2.2.3 Obtaining protein crystals

With a stock of highly pure protein (> 98% is desirable), whether isolated from its natural source or an expression system (e.g. bacterial, yeast, baculovirus) the initial crystallisation trials can begin. Ideally the protein sample should be fresh and mono-disperse i.e. the protein exists as a single oligomeric species e.g. monomer, dimer etc, and not as aggregates. The presence of a mono-disperse sample is best determined by *dynamic light scattering.* Always the protein should be correctly folded; this can be checked with some confidence by *circular dichromism* to ascertain secondary structure.

A number of techniques are available to promote crystal growth from a protein solution: *hanging drop, sitting drop, sandwich drop, dialysis buttons, gel,* and *microbatch* experiments. All techniques share the same underlying principle; to bring a protein solution to a supersaturated state to induce controlled precipitation and subsequent crystal growth . Protein will remain in solution up to a certain concentration dependent upon the individual properties of the particular protein and the characteristics of its solvent. When this concentration is exceeded the solution will no longer remain homogenous and a new state or phase will appear. For the crystallographer it is hoped that the protein enters a crystalline state. In vapour diffusion crystallisation techniques a supersaturated protein solution is reached by diffusion of water from a volume of protein in an aqueous solvent to a larger volume of the same solvent not containing protein, and then deliberately over-run to yielding protein crystals. Diffusion occurs spontaneously due to the difference in concentration of protein between the protein sample and the larger volume of solvent. Since this type of diffusion is by nature very slow the volume of the protein sample is reduced in a gradual manner and the coupled increase in protein concentration occurs in a similar manner (105, 106, 107, 108, 111). Coincidently, of course, the solvent constituents will also increase in concentration in the protein sample. The most common techniques used to grow protein crystals are the hanging (*Figure 2.2*) and sitting drop methods (92), largely due to their rather low-tech nature. Typically, a small volume (0.5 – 2 µl) of protein solution, usually still in its purification buffer, is added to an equal volume of aqueous solvent. The small volume of protein solution is then either suspended (hanging drop) or supported (sitting drop) over a well containing the same aqueous solvent as that mixed with the protein sample. A closed system is achieved by either sealing the set-up with either a glass cover slip or sealant tape. Spontaneous diffusion of water can then occur from the protein drop to the well solution.

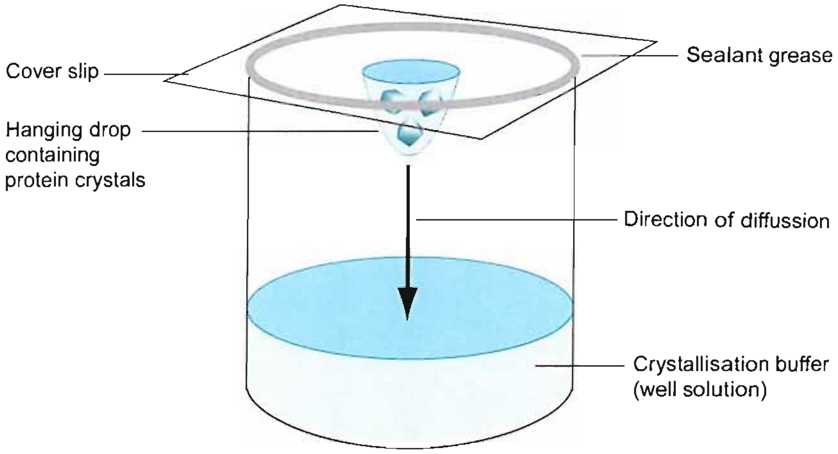

Cover slip
Sealant grease
Hanging drop containing protein crystals
Direction of diffussion
Crystallisation buffer (well solution)

Figure 2.2: Diagrammatic representation of the hanging drop vapour diffusion technique. *Vapour diffusion from the hanging drop to the well solution allows the concentration of the precipitant component and protein of the hanging drop to increase. In certain, usually highly specific, conditions this process causes the orderly assembly of protein molecules into an ordered lattice i.e. a protein crystal. Due to the small volume of the hanging drop it can be suspended from a glass cover slip, adhering due to its own surface tension.*

2.2.4 Crystal growth

Crystal growth occurs in three phases; nucleation, growth and growth cessation. Nucleation involves protein molecules of the supersaturated solution forming thermodynamically stable aggregates displaying a repeating arrangement. For macro crystal growth to occur from these aggregates, a critical volume has to be exceeded (93). If this critical volume is exceeded, the aggregate is termed a supercritical nucleus that is capable of sustaining further growth; if the critical volume is not met the aggregate will dissociate.

Nucleation is controlled by the level of super-saturation of the protein solution which is, in turn, controlled by protein solubility. At high concentrations the chance of protein molecules coming together to form a supercritical nucleus is obviously far higher than in a solution of low concentration. Therefore the higher the solubility of a protein in a given solvent the more likely nucleation will occur. The formation of the supercritical nucleus represents a high energy intermediate state: the energy barrier to crystal formation. The energy deficit is accounted for by generating the supersaturated protein solution with its associated high free energy. Once the energetically un-favourable process of nucleation has occurred crystal growth will proceed (111) (*Figure 2.3*).

Crystal growth is an energetically favourable process since protein molecules as part of a crystal will have a lower free energy (~ 5 kcal / mol) than protein molecules in solution (94). Crystal growth is driven by diffusion of protein molecules from solution to the solid state of the crystal and will continue at protein concentrations equal to, or slightly lower than those favourable to nucleation. To maximise crystal size only a few points of nucleation are desirable. As crystals grow there will be a concurrent decrease in soluble protein concentration and this will actually prevent further nucleation as the degree of super-saturation of the solution is decreased. The rate of crystal growth is then dependent upon the characteristics of the crystal; a poorly ordered crystal with its rough surfaces will grow at a faster rate than a well ordered crystal with smooth faces as a lower energy threshold exists for addition of protein molecules to rough surfaces than to smooth. In addition the shape of the crystal will influence growth rates; a flat crystal requires nucleation from two points as sheets of molecules grow one on top of the other; stepped faced crystals, that usually arise because of a screw axis symmetry operation, grow as columns and as such require a single point of nucleation and grow at a faster rate; finally, kinked faced crystals grow at the fastest rate as no further points of nucleation are required for crystal growth (105, 106, 107, 108, 111).

The final phase of crystal growth is the cessation of growth; this can occur due to a number of reasons. Most often cessation of growth is due to protein concentrations dropping so low as not to favour the relatively concentrated protein conditions required for crystal growth as the solid and solution phases reach equilibrium. In some cases crystal grow is limited by a phenomenon known as lattice strain where a certain protein crystal will only ever reach a finite size irrespective of protein concentration (95). In other cases poisoning of the growing face can occur where non-proteinous or damaged protein molecules are incorporated causing a defect in the crystal lattice (97).

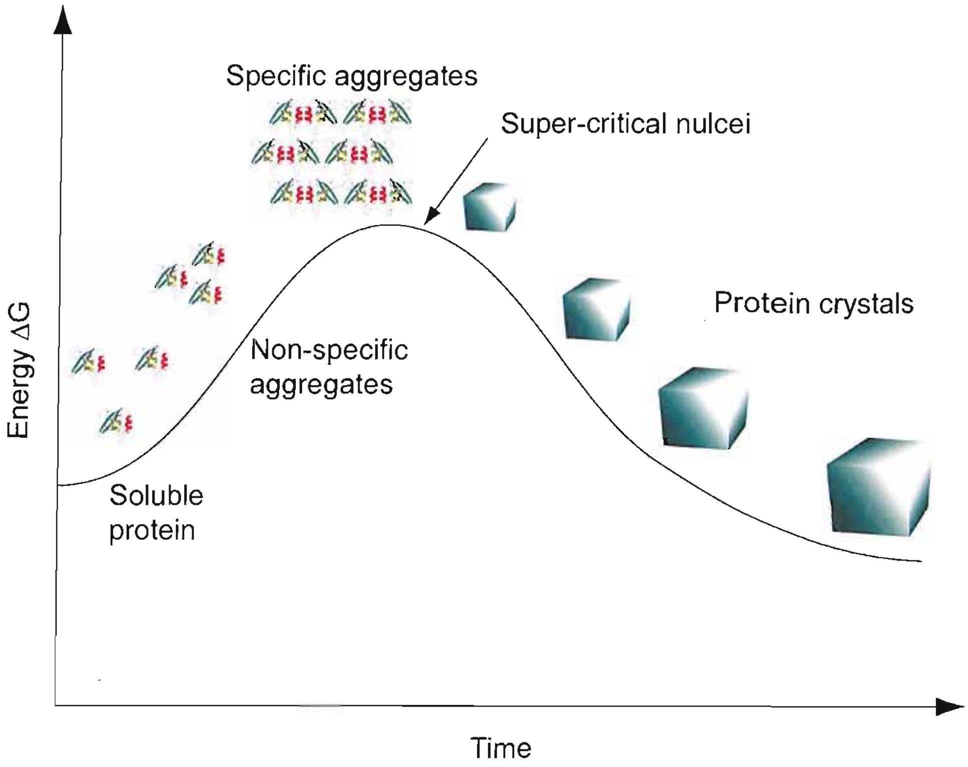


Figure 2.3: Protein crystallisation and crystal growth*.* *As with conventional chemical reactions, an energy barrier must be overcome in order for protein crystals to form. The supercritical nucleus represents the highest energy state of crystal formation. A protein solution in a supersaturated state possesses a relatively high free energy and can overcome this energy barrier leading to nucleation and subsequent crystal growth. As crystal growth continues the system becomes more ordered and free energy decreases.*

2.2.5 Factors affecting crystal growth

It follows that factors affecting protein solubility are vital to crystallisation. To affect the solubility of a protein it is the properties of its solvent in which it is in solution, that are altered to slowly approach and the necessary state of super-saturation. Such properties are: solvent pH, temperature, buffer type and concentration and the presence of additives (e.g. protein precipitants, divalent ions, detergents etc). Of course, protein concentration itself can be altered at the outset and routinely crystallisation trials are set-up over a range of concentrations since a protein may precipitate immediately in an uncontrolled manner in a particular solvent but at a lower concentration may reach a supersaturated state in a controlled manner to yield protein crystals. It is this very fine line between rapid uncontrolled protein precipitation and slow controlled nucleation and crystal growth that must be found. Unfortunately, finding this ideal solvent and its associated protein concentration is largely based upon experiments of trial and error that not only consume relatively large amounts of protein but can also be very time consuming. To aid in identification of the often elusive ideal condition to produce the desired large well ordered crystals, kits are commercially available that include typically up to 200 different conditions, with solvent characteristics as previously described being varied. Certain buffers, precipitants and organic solvents are known to have a relatively high success rate in promotion of crystal growth and screens usually devote a number of conditions to each of these solvents where the constituents are slightly varied. In this way, rather than wasting time and protein on certain conditions that only extremely rarely produce protein crystals, as with a *full factorial* approach (where all parameters of a matrix are sampled) the *sparse matrix* approach is often adopted where the matrix is skewed towards the more successful conditions. Statistically successful protein precipitants capable of yielding protein crystals suitable for X-ray crystallography use are; polyethyleneglycol (PEG), 2-methyl-2,4-pentanediol (MPD), along with some salts (predominantly NaCl and $(NH_4)_2SO_4$) and some alcohols. The concentrations of these precipitants along with buffer type and pH of the condition are varied to provide a comprehensive range of conditions whilst still providing a screen with a bias towards the most statistically successful conditions (111).

With the exception of PEG, each of the precipitants has the effect of dehydrating a protein solution by competing for water molecules and so a concurrent increase in protein concentration occurs bringing the solution to a supersaturated state. In a slightly different manner PEG can cause an increase in protein concentration by reducing the available solvent volume through the phenomenon of volume exclusion whereby the aqueous solvent undergoes a restructuring due to the presence of PEG (109), again causing a concurrent increase in protein concentration. Protein solubility is also affected by the ionic strength of its solvent, a weak ionic strength will favour protein aggregation but reduce protein solubility, a high ionic strength will, to a point, provide conditions that increase protein solubility thereby increasing the super-saturation of protein solution. In this way salts and alcohols can affect the dielectric constant of solutions and promote super-saturation and subsequent nucleation by increasing self association of protein molecules.

To better illustrate the process of nucleation and crystal growth a diagram describing the effect of protein saturation and solution properties is often referred to (*Figure 2.4*).

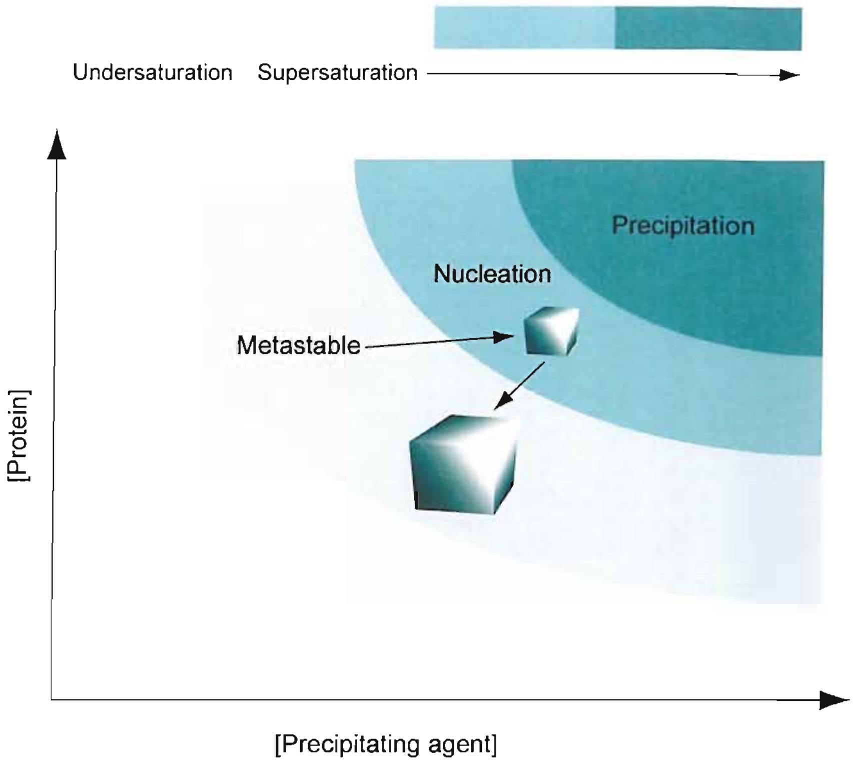


Figure 2.4: Nucleation and crystal growth as a function of concentration of protein and precipitating agent. *To promote nucleation and crystal growth, a super-saturated protein solution must be approached. The higher the super-saturation, the more likely nucleation will take place. To favour crystal growth therefore, the system must be forced to become as super-saturated as possible but without becoming so concentrated that uncontrolled protein precipitation occurs. By increasing the protein concentration and concentration of precipitating agent of the solvent the necessary super-saturation can be forced. Once nucleation has occurred soluble protein concentration will drop as protein molecules enter the solid phase represented by the crystal. This slightly lower level super-saturation favours crystal growth. Note: the 'metastable' state refers to a system not in equilibrium but able to persist for a relatively long time and defines the state between an under-saturated state and a super-saturated solution capable of nucleation* (105, 106, 107).

Initial indications of successful protein crystallisation are clean, sharp-edged protein crystals measuring at least 20 µm in all dimensions but preferable larger; 100 -200 µm. Poor quality crystals can sometimes be improved by addition of certain additives ("additive" screens are available to determine such additives), or by seeding (133), a technique where a crystal is crushed and the fragments used in a second vapour diffusion experiment to "seed" nucleation of protein molecules in the aim to improve crystal quality and size. It follows that a larger crystal will contain many more protein molecules than a smaller crystal and therefore will have a larger capacity to diffract X-rays and so provide higher quality data. However this is not always the case; a poorly ordered crystal irrespective of size will diffract X-rays poorly, and so the only real test of crystal viability is to test experimentally the diffraction on an X-ray beam.

2.3 Diffraction theory

2.3.1 Protein crystals scatter X-rays

It is the electron charge density surrounding the nuclei of atoms of a protein crystal that scatter incident X-rays. When an X-ray interacts with a point of charge it causes that charge to oscillate. A point of oscillating charge i.e. accelerating charge, acts as a new source of X-ray emission. The intensity of the X-ray emission is dependent upon the electron charge density at the point of X-ray incidence with the crystal. Whilst X-rays are capable of interacting with atomic nuclei, due to the relatively massive size of the nucleus, oscillations induced by collision of X-ray are so small as to be negligible, and so therefore is a nucleus's ability to diffract X-rays in a manner of use to the crystallographer. It is therefore the aim of the crystallographer to map electron density rather than directly determine nuclei position (105, 106, 107).

An X-ray beam of width 0.1 mm^2 impinging upon a crystal of average size i.e. 50 µm^2 will cause multiple points of charge to oscillate (*Figure 2.5*). A single scattered X-ray will therefore be prone to interference from scattered X-rays originating from other points of charge within the crystal. Due to such interference, the intensity of all scattered X-rays, irrespective of their point of incidence with the crystal or direction of scattering, will depend upon the charge distribution throughout the whole illuminated sample; in the case of the 50 µm^2 crystal electron charge density spanning the entire crystal will contribute to the intensity of each of the diffracted X-rays since the incident beam size 0.1 mm^2 is larger than the crystal itself.

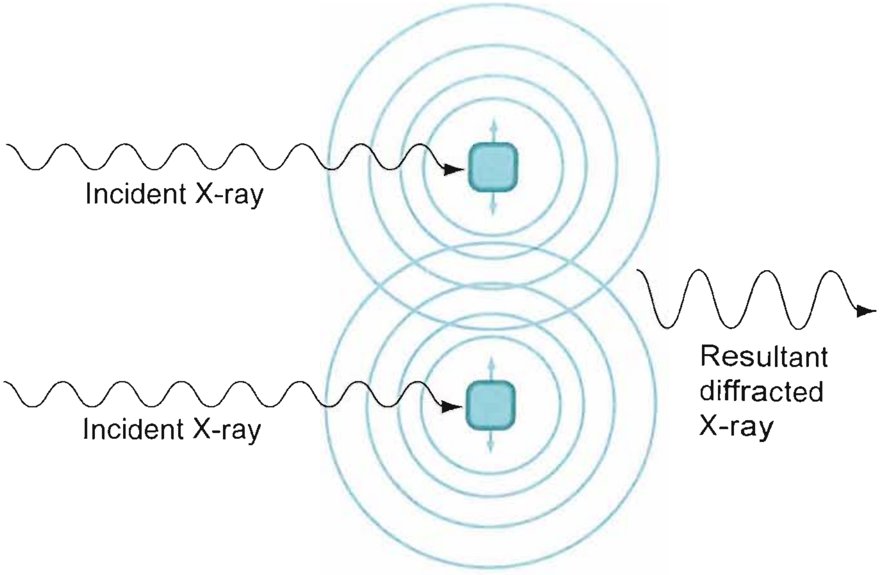


Figure 2.5: As an X-ray interacts with a point of charge it will cause that charge to oscillate*. Two oscillating points of charge will emit electromagnetic radiation of the same wavelength as the impinging X-ray. As the emerging electromagnetic radiation has the same properties as the impinging X-ray we consider the original X-ray to have been diffracted. The diffracted X-rays will interfere with each other; therefore point 1 contributes to the intensity of the diffracted X-ray from point 2 and vice-versa.*

The simplified situation described in *Figure 2.5* can be extended to a protein crystal consisting of numerous points of charge, each diffracting X-rays that will interfere with each other. However, only some interference will be constructive and result in a diffracted X-ray; most diffracted X-rays interfere de-constructively.

An X-ray emerging as a result of constructive interference will be a complex wave i.e. its waveform is the result of diffraction from multiple points of charge and therefore carries information about each point. The mathematical function, Fourier transformation, allows the waveform to be broken down in to the constituent sinusoids that comprise the diffracted X-ray (105, 106, 107).

2.3.2 Diffraction from Bragg planes

In the Bragg model of diffraction, X-rays are considered to be diffracted from sets of imaginary planes that dissect the unit cell. These planes are sometimes referred to as *Bragg planes.*

For two incident X-rays of identical wavelength to be scattered and interfere constructively they must emerge in phase i.e. at the same point in the wave cycle (possess the same phase angle) (*Figure 2.6*). It follows that for two diffracted X-rays to be in phase the difference in distance travelled (path-length) by two X-rays must be equal to an integral value of their wavelength. By simple trigonometry we can demonstrate the difference in path-length must be 2dsinθ; this is the essence of *Bragg's Law.*

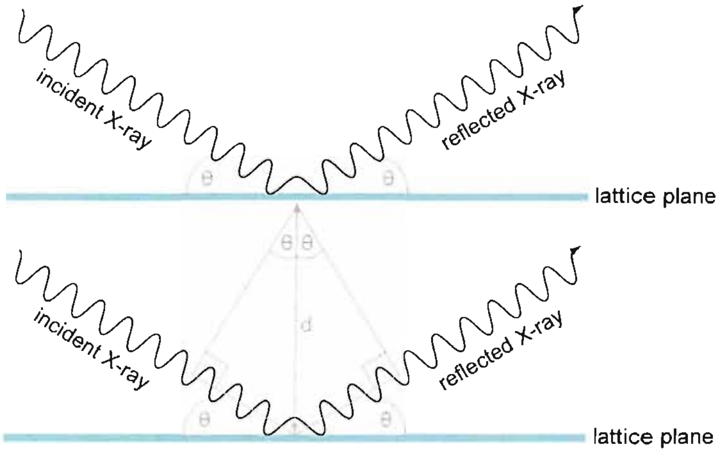


Figure 2.6: Satisfaction of Braggs Law. *The angle of diffraction of an X-ray is equal to its angle of incidence; call this angle θ. If two diffracted X-rays, diffracted from electron charge in two different points of protein molecule, are to interfere constructively the distance between the points of X-ray incidence must be equal to an integral value of the incident X-ray beams wavelength; call this distance d.*

In summary, Bragg planes can be considered as sources of X-ray scattering where diffracted X-rays emerge in phase and therefore interfere constructively. According to Bragg's law, such planes must be separated by a distance equal to an integral multiple value of the wavelength of the incident X-ray (105, 106, 107).

2.3.3 Identifying Bragg planes

If a hypothetical two dimensional crystal lattice is considered we can draw sets of planes (Bragg planes) each separated by an integral value of the impinging X-ray beam wavelength (*Figure 2.7*).

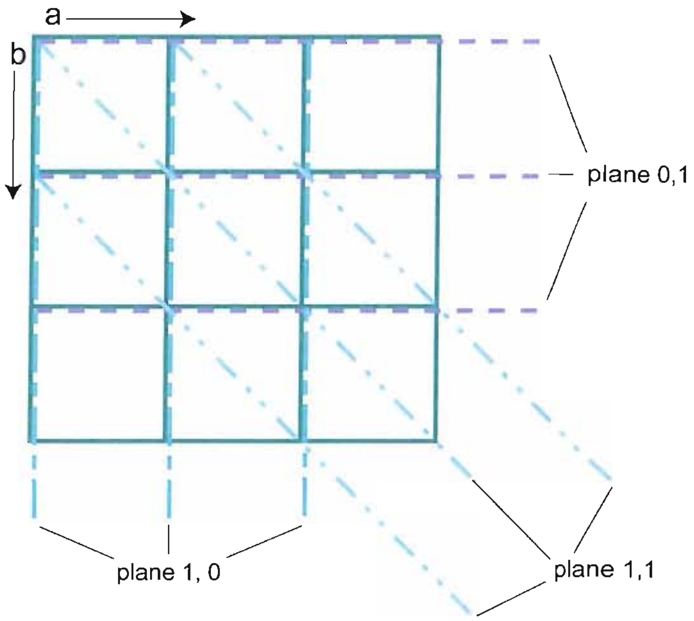


Figure 2.7: Bragg planes in two dimensions. *Shown is a hypothetical two dimensional crystal lattice, comprising nine unit cells. A set of Bragg planes can be drawn that are parallel to the cell edge labelled 'b' and separated by a distance equal to the cell edge 'a'. Such a set of planes can be given the index 1, 0; corresponding to the number of times the planes cross the 'a' and 'b' cell edges per unit cell. A further set of planes can be drawn that are parallel to the 'b' cell edge and are separated by a distance equal to the cell edge 'a' , these planes can be called 0,1. Similarly we can draw a set of planes that cross both the cell edges 'a' and 'b' once per unit cell (i.e. pass through diagonally opposed lattice points); this set of planes would be given the index 1,1.*

Extending this situation to a three dimensional crystal lattice (*Figure 2.8*) where we have cell edges 'a', 'b' and 'c', a set of planes that crosses each cell edge once could be drawn and these would be given the index 1, 1, 1.

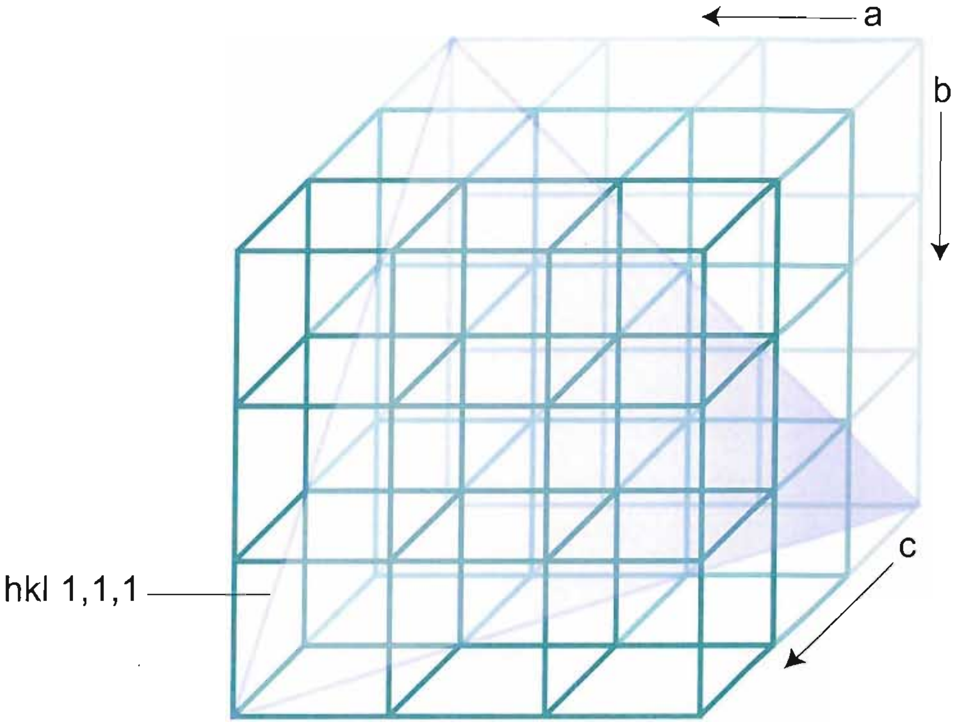


Figure 2.8: Bragg planes in three dimensions. *A three dimensional lattice with the plane 1,1,1 shown passing through lattice points of the a, b, and c lattice axis. Only one plane of the 1,1,1 set of planes is shown for clarity of the diagram.*

So rather than referring to the planes as fractions of the '*a*', '*b*' and '*c*' cell edges they are referred to as the *h*, *k*, *l* planes where it is assumed that a value of *h* of 1 equals *a*/1, a value of h of 2 equals a/2; and similarly for *k* and *l*. It follows that the distance, d_{hkl} between a set of Bragg planes with the index 1, 1, 1 must be larger than a set of planes 2, 2, 2. Further, the closer Bragg planes are together, the steeper the angle of incidence and, therefore, the diffraction of an X-ray must be to satisfy Braggs Law. Finally, and most importantly to the crystallographer, the information carried by X-rays diffracted from a set of Bragg planes with a high index is greater than diffraction from planes with a low index. The higher the index, the higher the resolution of the information the diffracted X-ray will carry since it has been diffracted from planes closer in space.

In summary, adopting the Bragg model of diffraction, the manifestation of diffraction by a single set of Bragg planes is a single diffracted X-ray of complex waveform. The complexity of the diffracted X-ray is as a result of interference of diffracted X-rays carrying information pertaining to electron density spanning the entire unit cell (105, 106, 107).

2.3.4 From diffracted X-rays to electron density maps

To map electron density within the unit cell, the crystallographer requires a means of “focusing” diffracted X-rays. Currently the greatest detail an X-ray focussing lens can resolve are points separated by 1 µm; a distance several orders of magnitude too large to resolve detail of a protein’s substructure. However if Bragg reflections are collected by means of an X-ray detector subsequent computational reconstruction of the unit cell electron density can be made, though not directly. The only property of a diffracted X-ray that can be measured is its intensity as it arrives at the X-ray detector. This presents a problem to the crystallographer as it would be desirable to also record the point in the wave cycle of the diffracted X-ray at which it arrives at the detector i.e. the phase of the diffracted X-ray – this is the much bemoaned ‘phase problem’. To overcome the ‘phase problem’ the crystallographer has to use indirect methods.

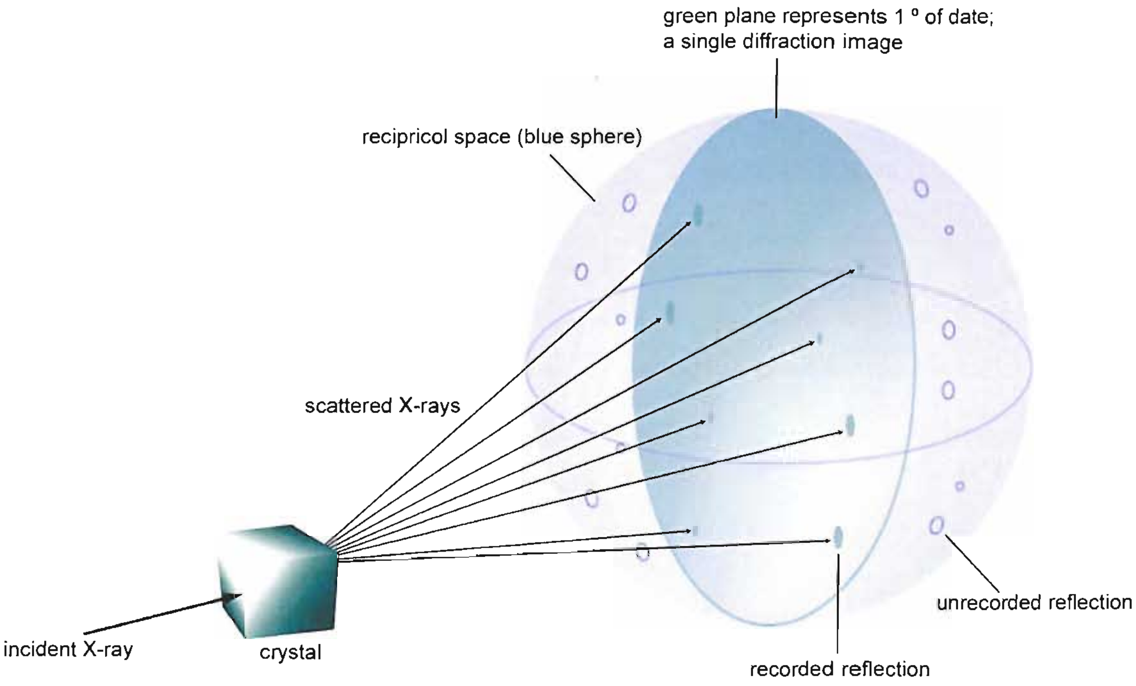
A diffraction image comprises many reflections (typically in the thousands) each corresponding to diffraction of X-rays from a single set of Bragg planes whilst the crystal is placed in the X-ray beam in a single orientation. In practice, the crystal is rotated through a small angle, typically 1°, during recording of diffraction to generate a single diffraction image. Reflections closer to the centre of the image correspond to X-rays of a lower angle of diffraction (θ) and therefore from widely spaced planes carrying only low resolution detail. Reflections towards the outside of the diffraction image result from X-rays diffracted to a higher angle and so from planes closely spaced and carrying high resolution detail. The number and distribution of reflections on a single diffraction image is dependent upon the symmetry of the protein crystal and as such it describes nothing of the electron density within the unit cell; this information is contained within the intensity of the reflections themselves. Reflection intensity varies widely between reflections and is dependent upon the electron density that lays along the plane that gave rise to the reflection. A strong reflection will be generated from diffraction occurring from planes along which lie a large amount of electron density; in contrast, a weaker reflection would be seen from planes along which much solvent lie as a result of the weak diffraction of X-ray by solvent ions.

Even in the simplest case of no symmetry in the protein crystal (i.e. a P1 lattice) the diffraction image will still include reflections arising from diffraction from either side of a Bragg plane. Such reflections are known as Friedel pairs. Friedel's law stipulates that since such reflections are generated by diffraction from the same set of planes they must have identical intensity. In practice, Friedel pairs displaying equal intensity are difficult to identify easily on a single diffraction image, although Friedel pairs do become far more apparent when trans-sections through *reciprocal space* are viewed (105, 106, 107).

2.3.5 X-Rays in reciprocal space

Reciprocal space is a theoretical system that is, in effect, a three dimensional reconstruction of all the two dimensional diffraction images of a single dataset. *Reciprocal space* therefore describes the locations and intensities of all recorded *h*, *k*, *l* reflections. Each *h*, *k*, *l* reflection in *reciprocal space* is termed a *reciprocal lattice point.* Reciprocal space exists at the point before 'focussing' of the X-rays, so correspondingly all cell dimensions and co-ordinates in *reciprocal space* are the inverse of those in *real space*. Although it must be appreciated that each *h*, *k*, *l* index does not have a directly related *x*, *y*, *z* co-ordinate in real space, since the reflection at an *h*, *k*, *l* index is as a result of diffraction from all points in the illuminated crystal (*Figure 2.9*).

Considering reflections in *reciprocal space* helps in understanding the relationship between index and the spacing of Bragg planes giving rise to the reflection. As stated a reflection with a high index is generated by closely spaced planes. If the index is considered as a vector 's', by simple geometry we can prove that the higher the index the larger the resultant vector. Since reflections exist within *reciprocal space* and reciprocal space is an inverse representation of real space, it follows that s = 1 / d_{hkl} therefore the higher the index the smaller is d_{hkl}, and closer are the planes giving rise to the reflection. This returns to the concept that higher index reflections carry higher resolution information as regards to protein substructure (105, 106, 107).



green plane represents 1 ° of date; a single diffraction image
recipricol space (blue sphere)
scattered X-rays
incident X-ray
crystal
unrecorded reflection
recorded reflection

***Figure 2.9: Reciprocal space**. Reciprocal system that is concerned with the distribution and intensity of X-rays diffracted from all Bragg planes of the protein crystal. In reciprocal space all h, k, l intensities have been scaled to take account of fluctuations in X-ray diffraction due to practical restraints discussed previously.*

2.3.6 The Ewald construct

As previously discussed, in a single crystal orientation in the X-ray beam diffraction cannot be recorded from all diffraction planes so it is necessary to take images at a succession of angels by rotation of the crystal. The number and range of angles is dependent upon the space group of the protein crystal and the associated level of symmetry.

For crystals of higher symmetry a lower number of images through a smaller range of angles will need to be taken. To further explain the need for the rotation method of data collection a model known as the *Ewald construct* may be considered (*Figure 2.10*). Essentially a simple extension of Bragg's law applied to reciprocal space.

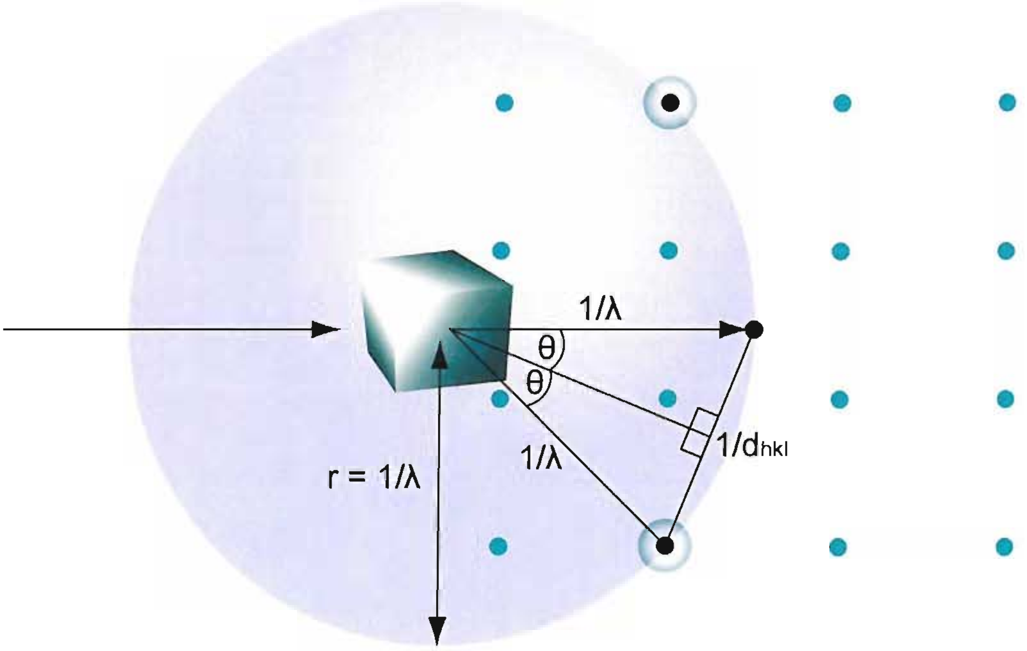


Figure 2.10: The Ewald construct. *Since the Ewald construct (also termed the Ewald Sphere) exists in reciprocal space, if a sphere of radius equal to the inverse of the wavelength of the incident X-ray beam is drawn, any reciprocal lattice points that touch the surface of that sphere will satisfy Bragg's law and a diffracted X-ray will result. The origin of the reciprocal lattice is drawn at the point where a non-diffracted X-ray passing directly through the crystal exits the Ewald sphere, indicated here as the black coloured dot and given the index 0, 0, 0. At some point during rotation of the protein crystal in the X-ray beam, every reciprocal lattice point (light green dot) will coincide with the surface of the Ewald sphere; Bragg's law will be satisfied and the resultant reflection can be recorded (dark green highlighted dots).*

At some point in rotation of the crystal all reciprocal lattice points will come into contact with the surface of the *Ewald sphere* and therefore satisfy Bragg's law and the resultant constructive interference will generate a diffracted X-ray. As touched on previously, such X-rays will produce a single reflection (spot) on the diffraction image with each assigned an index corresponding to the *h, k, l* plane (or Bragg plane) generating them. The first reflection from the *reciprocal lattice* origin on the *h* axis corresponds to diffraction of X-rays from the *h, k, l* plane 1, 0, 0; the distance between the reciprocal origin and this reflection is therefore equal to the inverse of the length of the real cell edge *a (Figure 2.10)*. The same is true for the *k* and *l* reciprocal axis and *b* and *c* real cell edges respectively. Further, since the spacing of reciprocal lattice points is the inverse of the Bragg planes in real space - lower index reflections occupy space closer to the reciprocal lattice origin, higher index reflections occupy space further from the origin.

Whilst Bragg's law does provide a convenient way of describing X-ray diffraction by a protein crystal and why reflections appear in the diffraction image with the distribution they do, in reality such tangible planes as sources of X-ray diffraction do not exist. To understand how structural information can be extracted from diffraction data an alternative model must be considered where diffraction can be thought of as occurring from each atom in the illuminated portion of the crystal (105, 106, 107).

2.3.7 An alternative model of diffraction

The diffracted X-ray giving rise to a single reflection will have a waveform dictated by interference of X-rays scattered by all atoms in the unit cell. Further, the waveform will no longer be sinusoidal as with the incident X-rays but due to summation of multiple diffracted X-rays the waveform becomes complex in nature. The mathematical description of a complex diffracted X-ray giving rise to a single reflection is called a *structure factor*; F_{hkl}, (*eq.1*) of amplitude F arising from Bragg planes *h, k, l* in the Bragg model of diffraction.

$$F_{hkl} = \Sigma\, f_j e^{2\pi i(hx_j + ky_j + lz_j)} \qquad (eq.1)$$

Such a mathematical description constitutes a *Fourier series* that comprises individual terms each describing the contribution of a single atom to the structure factor. Individual atoms can be considered as spheres of electron density where the scattering attributed to a single atom, *j*, depends upon the nature of that atom, f_j,(the atomic scattering factor) and its position in the unit cell; x_j, y_j, z_j; the position defining the phase angle of the diffracted X-ray.

Possibly a more useful description of a structure factor is to consider it as a sum of diffraction of X-rays from small volumes of electron charge (*eq.2*) in the unit cell rather than from individual atoms; where electron charge can be described as an average density $\rho(x, y, z)$ over a volume element. In other words; an integral term covering the unit cell volume, V, is comprised of the contributory diffraction from small volumes of electron charge at multiple locations in the unit cell.

$$F_{hkl} = \int_V \rho\,(x, y, z)\; e^{2\pi i(hx + ky + lz)} dV \qquad (eq.2)$$

By application of Fourier methods, a complex wave can be de-convoluted into its simple constituent waves each possessing an amplitude and phase angle; revealing the density of the point of electron charge causing the diffraction and the distance between that point of charge and the X-ray detector allowing its positioning in space (105, 106, 107).

2.3.8 Structure factors as vectors

It is useful to think of structure factors as consisting of vectors; each vector possessing an amplitude and phase angle and describing a simple diffracted X-ray as a result of scattering by a single atom or volume element. A useful visual aid to describing structure factors as vectors is an Argand diagram (*Figure 2.11*) where all the vectors comprising a single structure factor are placed nose to tail allowing the final vector describing the structure factor to be easily calculated i.e. its amplitude and phase angle, Φ.

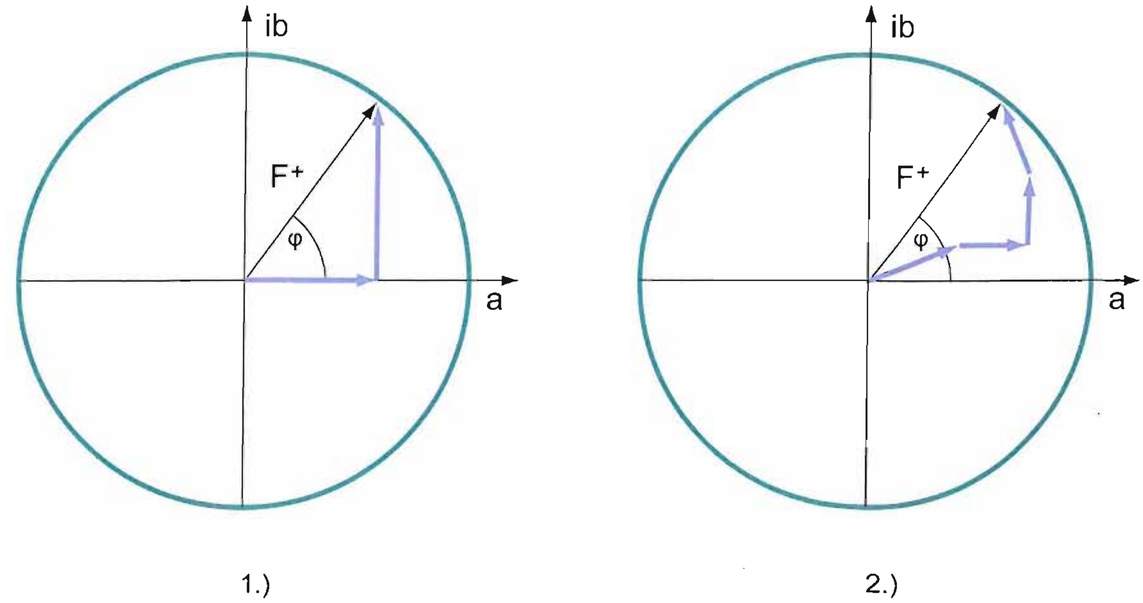


***Figure 2.11: An Argand diagram**. the contributions by individual atoms or volume elements to the resultant structure factor; F_{hkl} (black arrow), are shown as vectors (blue arrows).*

It is the reverse of this situation that faces the crystallographer; the amplitude of the structure factor vector is easily calculable by taking the square root of the intensity of the relevant Bragg reflection at the X-ray detector. However, the structure factor phase angle can not be directly measured from a diffraction data set of a native protein; though, if relatively difficult to calculate, can be determined using indirect techniques. With the structure factor vector fully described; amplitude and phase angle, it is its Fourier transform into its constituent vectors that yields details of protein substructure along a particular plane as described by the Bragg indices of the reflection.

In a similar way electron charge density lying along all Bragg planes in the real cell can be expressed as a Fourier series that consists of all values of recorded F_{hk} (*eq*.3)$_{l}$.

$$\rho_{(x,\,y,\,z)} = 1 / V \sum_{h}\sum_{k}\sum_{l} F_{hkl}\, e^{-2\pi i\,(hx + ky + lz)} \qquad (eq.3)$$

So the periodic function describing the entire unit cell electron charge density becomes a Fourier transform of all recorded structure factors in reciprocal space. However, as with the Fourier series describing a single structure factor; to satisfy the electron density equation the frequencies, amplitudes and phases of all structure factors must be known. The frequency of the complex wave resulting in each structure factor is that of the incident X-rays, the amplitude in the square root of intensity, and the phase angle, is sought by indirect techniques (105, 106, 107).

2.4 Phase angle determination

Much of the effort required to solve a protein structure by X-ray crystallography is expended upon determining the phase angle of the diffracted X-rays. Unfortunately for the crystallographer, X-ray detectors of the type used in protein crystallography are insensitive to the phase angle of an X-ray incident upon their surface. For a detector to record the relative phase angles of incident X-rays it would need to be capable of taking timed measurements at fractions of the period (time for one complete wave cycle) of the X-ray. Considering that the frequency of X-rays of the longest period used in protein crystallography is approximately 2 x 10^{18} wave cycles per second, if a detector were to make a precise enough measurement of phase angle, at least 1/5th of a wave cycle, measurements every 10^{-19} seconds would have to made. This is way beyond the current capabilities of any present detector technology. Hence at this time an indirect method to phase determination has to be taken (105, 106, 107).

Before discussing the techniques available for phase determination it is necessary to go back to the concept of describing structure factors as vectors, but in this instance phase-less structure factors as phase-less vectors. This is known as the calculation of a *Patterson function.*

2.4.1 The Patterson function

The *Patterson function* (139) is essentially the Fourier transform of the structure factor amplitudes. By definition, the Patterson function does not rely upon any structure factor phase information and can therefore be calculated directly from the measured amplitudes of the structure factors. This 'phase-less' Fourier transform is equivalent to a map of the protein structure's inter-atomic vectors.

Consider a protein molecule comprising of just three atoms *a*, *b* and *c*; vectors, denoted by vector length, can be drawn from each atom to the other two atoms resulting in a total of six inter-atomic vectors; *ab*, *ac*, *ba*, *bc*, *ca* and *cb* (*Figure 2.12*). Each vector consists simply of an amplitude. A *Patterson map* can be constructed where all six inter-atomic vectors are drawn from an origin named 0, 0, 0 to produce six peaks on the map (*Figure 2.12*). To be precise, such a Patterson map would consist of nine peaks; six relating to inter-atomic vectors as described and three corresponding to vectors between each atom and itself to produce a peak at the origin where these three vectors, of zero amplitude, overlap.

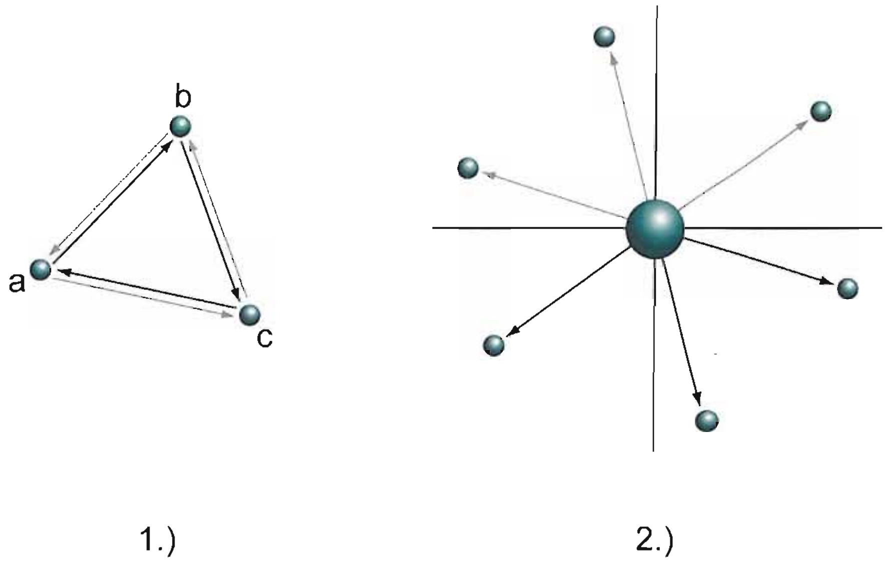


Figure 2.12 Inter atomic distances described as vectors. *Six vectors describing three atoms a, b, and c separated in three dimensional space can be calculated from diffraction data producing the resultant three-dimensional Patterson map.*

In practice, Patterson maps consist of many more peaks and for a molecule of N atoms there will be N^2 peaks. A number of peaks equal to N will exist at the origin of the Patterson map therefore the total number of non-origin inter-atomic peaks will equal $N^2 - N$. Consider an average sized protein of 200 residues; each residue of an average 10 atoms will result in nearly 4 million peaks! In addition, since a complete set of *structure factors* describes the whole unit cell, not simply a single protein molecule, in practice Patterson maps include peaks corresponding to inter-molecular vectors as well. Not surprisingly, their calculation and interpretation is a computer based procedure. It should be appreciated that the inter-atomic vectors of a Patterson map are representative of real space atomic separations and therefore the Patterson function exists in real space; after all, the Patterson function is the Fourier transform of phase-less *structure factors* that exist in *reciprocal space*, therefore the *Patterson function* must exist in *real space* (105, 106, 107).

It is the extreme complexity of Patterson maps of macromolecules that prevents their solution directly. Unfortunately, due to the massive number of permutations of possible atomic arrangements, directly solving the spatial locations of atoms in the unit cell from Patterson maps is only applicable in solution of structures consisting of only a few atoms. Due to computational restrictions, this approach is not currently suitable for macromolecular complexes. However, as discussed later, obtaining a structural solution for a small number of atoms using just Patterson maps is key in locating heavy atoms during *SAD* and *MAD* phasing.

2.4.2 Phase determination techniques

There exist three main techniques for phase determination; *molecular replacement* (MR), *multiple isomorphous replacement* (MIR) and *multiwavelength / single wavelength anomalous dispersion* (MAD/SAD). Each of these techniques provides an initial estimate of phases that, when combined with the experimentally determined *structure factor amplitudes,* provide an interpretable map of electron charge density within the *unit cell.* Through iterative rounds of structural modelling (manual fitting of a model of the target protein to the experimental data i.e. the electron density) and automated refinement; computer based adjustment of the structural model to calculate modified phases, the initial phase estimates can be improved.

As this thesis mainly concerns the technique of phasing by *MAD*; the techniques of *MR* will only briefly be discussed with *MIR* omitted entirely due to its similarity in basic principles to *SAD* or *MAD* phasing

2.4.3 Molecular replacement

Structure factors, and hence phases, can be calculated from a previously solved protein structure, the *model*, and applied to a set phase-less *structure factor amplitudes* of the unsolved protein, the *target*, to produce an initial electron density map. For molecular replacement to be successful a relatively high degree of structural homology between the model and target proteins has to exist (98). Sequence homology is a reasonable indication of structural homology; if the sequence homology of the putative model and target protein is < 25%, molecular replacement is unlikely to work. In practice, a sequence homology of approximately 50% is usually required. Also, it is important to have a dataset with a completeness of close to 100% i.e. near 100% of all predicted reflections at a certain resolution have been recorded during data collection.

The model structure is placed in the target structure unit cell and moved until it is superimposed over the target protein. A rotation matrix [C], and translation vector $\underline{d}$ define the orientation and position of the search model when superimposed over the target structure. The transformation between the search model and target protein where X represents a matrix defining position vectors of the search model structure and X' a matrix defining position vectors of the target structure can be expressed as in *eq.4*.

$$\underline{X}' = [C]\underline{X} + \underline{d} \qquad (eq.4)$$

To superimpose the search model over the target protein correctly, six parameters must be defined; three *rotation* parameters corresponding to rotation of the search model calculated as a function of Eurlerian angles; α, β, and γ; and three *translation* parameters corresponding to the necessary translations along the x, y and z axes. To simplify matters and reduce the computational load, MR is usually completed in two stages; the *rotation search* and the *translation search* (134).

Most programmes used in determination of the rotation search matrix rely upon Patterson methods (100). The Patterson functions for both search model and target protein are calculated for *self-vectors* i.e. intra-molecular vectors and are then subject to the rotation search, where the *Patterson self-vectors* are compared at different orientations of the *search model.* It is desirable to conduct the rotation search over a restricted shell with inner and outer radii limits in Patterson space to eliminate the origin peak, and cross-vectors; that is inter-molecular vectors of symmetry related protein molecules. Since all vectors of the Patterson map are shifted to the origin there is no need for any translation for superimposition during the rotation search. Agreement between the search model and target structure Patterson following the *rotation search* can be expressed as; R, calculated at different values of the rotation matrix [C] as described by *eq.5.*

$$\mathrm{R} = \int P_T(\underline{u})\ P_S([\mathrm{C}]\underline{u})\ d\underline{u} \qquad (eq.5)$$

Where the $P_T(\underline{u})$ is the Patterson of the target protein at position vector $\underline{u}$ and $P_S([\mathrm{C}]\underline{u})$ the Patterson of the search model rotated by matrix [C].

The translation search matrix can also be determined by Patterson methods although, in contrast to the rotation search, only the overlap of cross-vectors is measured. The rotation search is usually completed before the translation search, so that self-vectors can be eliminated from the translation search on the basis that search model and target protein self-vectors must overlap for a good solution; non-overlapping vectors can therefore be identified as cross vectors. The shift vector $\underline{d}$ defines the position of the search model so that it overlaps the target protein. A measure of overlap of cross-vectors of the search model and target structure is expressed as; *T*, calculated at different position vectors $\underline{t}$; where $P_c(\underline{u}, \underline{t})$ defines the search model Patterson cross-vectors and $P_o(\underline{u})$ the Patterson function of the target structure by *eq.6.*

$$T(\underline{t}) = \int P_c(\underline{u}, \underline{t}) P_o(\underline{u}) d\underline{u} \qquad (eq.6)$$

A good solution for the rotation search is passed into the translation search and if promising peaks are seen in both functions, phases from the model structure may be applied to the target protein data to obtain interpretable electron density maps, and assuming the sequence is known, to place the target protein into density. Since phases are obtained from a previously solved structure MR is not considered a technique where phase information is determined experimentally (105, 106, 107).

2.4.4 SAD / MAD phasing

2.4.4.1 Overview

The techniques of experimental phasing by *single-wavelength anomalous dispersion (SAD)* and *multi-wavelength anomalous dispersion (MAD)* rely upon the phenomenon of *anomalous scattering* by heavy atoms within a protein crystal. Anomalous scattering arises when the energy of an incident X-ray on an atom is equivalent to the transition energy of an outer shell electron, resulting in promotion of the electron to a higher energy orbital and the atom to an excited state. The atom will momentarily exist in this excited state before the outer shell electron returns to its original orbital and in doing so emits electro-magnetic energy as an X-ray but of altered phase to the incident X-ray that originally led to the promotion of the electron. The altered phase of the emitted X-ray leads to the break down of Friedel's law that states that reflections as a result of X-ray diffraction from both sides of a single set of Bragg planes, F_{hkl} and F_{-h-k-l}, belong to a pair of reflections (Friedel pair) of identical magnitude and opposite phase. During SAD/MAD phasing experiments the anomalous component of scattering provided by heavy atoms causes reflections of a Friedel pair to have *different* magnitudes, measurable as differences in intensity. It is these differences that can be exploited to spatially locate heavy atoms in the unit cell. Heavy atom phases can subsequently be located that can then be extended to protein atoms i.e. all atoms other than the heavy atoms of the protein (105, 106, 107).

Intrinsic to phasing by SAD or MAD is, of course, the presence of heavy atoms within the protein; that is atoms of atomic mass from iron (A_m=26), to palladium (A_m=46). Such atoms possess *absorption edges*; a transition energy corresponding to the energies of X-rays used during macromolecular X-ray crystallography (126). Whilst it is possible to promote C, N, and O atoms to excited states their absorption edges are at energies of X-rays not used in macromolecular crystallography. For proteins that do not contain a functional heavy atom e.g. iron co-ordinating the *haem* groups of *haemoglobin*, heavy atoms may be introduced to a protein crystal by two methods; *heavy atom soaks* (137), similar to those used in *MIR* phasing or the incorporation of a heavy atom during recombinant expression of the target protein, often in the form of *selenomethionine*; a selenium containing derivative of methionine (127). Both methods are equally as successful and the decision as to which to use is dependent upon the situation; for instance it would be impossible to prepare a selenomethionine derivative of a protein obtained from its natural source Further, selenomethione derivatives can only be prepared for proteins recombinantly expressed in bacterial systems (105, 106, 107).

With advances in computational software, the ability to locate ever increasing number of heavy atoms and phase ever larger proteins has resulted in SAD/MAD phasing becoming an increasingly routine method of macromolecular structural solution. The key theoretical elements of phasing by SAD/MAD will now be discussed in more detail.

2.4.4.2 X-Ray absorption near absorption edge energies

Also termed anomalous dispersion, anomalous scattering describes the scattering behaviour of X-rays by heavy atoms when the energy of the incident X-ray is equivalent to the energy required to promote an outer shell electron to a higher energy orbital. This energy is known as the absorption edge of the heavy atom as it the absorption of the incident energy that excites the outer shell electron. The excited electron will only exist in this excited state momentarily then fall back to its original energy level. In doing so the electron must lose the energy it has gained that promoted it to the higher energy state. The electromagnetic energy emitted by the regressing electron is identical in amplitude and wavelength to the incident X-ray that caused the excitation but has a phase-lag; that is its phase is altered in respect of the incident X-ray. Fortunately, the phase is always shifted by 90° compared to the phase of the incident X-ray irrespective of the type of heavy atom or protein environment. Since all Bragg reflections are as a result of interference of all diffracted X-rays of all diffracting atoms, including heavy atoms, in the protein molecule, each reflection, and therefore structure factor, will possess an anomalous component. The manifestation of the anomalous component is a difference in the recorded intensity between the reflections of Freidel pairs. To demonstrate the phenomenon of X-ray absorption and resulting anomalous scattering we first need to consider "normal scattering" i.e. X-ray diffraction away from absorption edge energies where the anomalous component is zero.

2.4.4.3 Normal scattering

As previously described, an X-ray incident upon an atom will cause the charge surrounding the atom to oscillate. The point of oscillating charge is itself a source of X-ray emission where the X-ray emitted is of identical wavelength and amplitude of the incident X-ray. We usually consider the emitted X-ray to possess the same phase angle as the incident X-ray, however, if we are being strict the emitted X-ray is 180° out of phase with the incident, although since this phase shift is identical for all atoms it is defined as a zero phase shift for scattered X-rays.

Normal scattering gives rise to reflections of Friedel pairs having identical intensities but opposite phase i.e. Friedels law is obeyed: $F_{hkl} = F_{-h-k-l}$ (*Figure 2.13*).

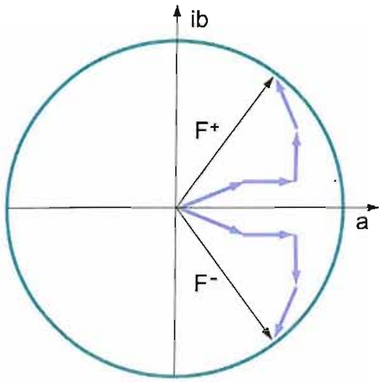


Figure 2.13: Normal X-ray scattering. *Structure factors F^+ and F^- belong to a Friedel pair of reflections and have arisen from normal scattering from a single set of Bragg planes. Despite opposite phases the resultant intensities of reflections F^+ and F^- are identical and obey Friedel's law. The vectors F^+ and F^- comprise of identical contributions from protein atoms (though greatly simplified here) so F^+ and F^- are therefore of identical length (and therefore amplitude) $F_{hkl} = F_{-h-k-l}$, but opposite phase angles; $\varphi_{hkl} = -\varphi_{-h-k-l}$* (105, 106, 107).

2.4.4.4 Breaking Friedel's law during anomalous scattering

When anomalous scattering occurs (at X-ray energies close to the absorption edge energy) the contribution to each reflection by the anomalously scattering atom is modified resulting in reflections, and therefore structure factors, of different intensities and amplitudes, respectively, to those at energies where the anomalous component is zero. If we again consider structure factors as vectors, two corrections to each structure factor are made - f' that modifies the length of the vector and f" modifies its phase angle.

If we again consider a vector diagram as a means of describing reflections of a Friedel pair and the resultant structure factors F_{PH}^{+} and F_{PH}^{-}, we see their magnitudes are no longer equal as the f" component always possesses a *positive* value - Friedel's law is broken (*Figure 2.14*). Such differences are termed *Bijvoet differences* (110) and the measurement of these is central to phasing by SAD/MAD.

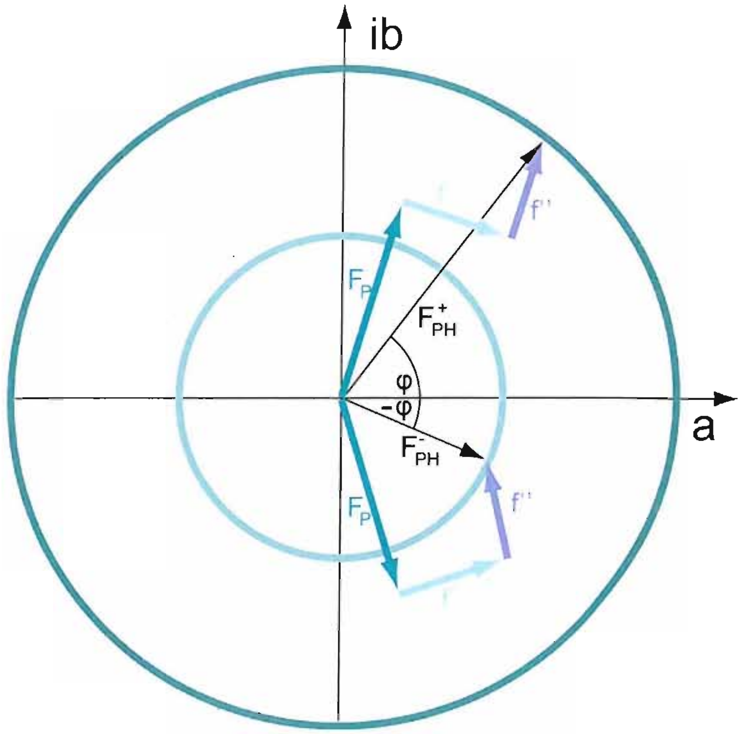


Figure 2.14: Anomalous scattering*. The vector description of structure factors resulting from Friedel paired reflections at energies equal (or very close to) the absorption edge energy of a heavy atom. F_P represents a structure factor where there is no anomalous contribution; F_{PH} represents a structure factor of the same indices but where there is an anomalous contribution (F_P is modified to F_{PH} when anomalous scattering occurs). The contribution of f" is always positive and holds clear ramifications for the magnitude and phase of the resultant vectors F_{PH}^+ and F_{PH}^- (Bijvoet differences). Dispersive differences (between datasets; see following paragraph) can be described in a similar way but with the vector F_P renamed as $F_{\lambda 1}$ representing a structure factor at a wavelength where f" in zero; and F_{PH} renamed $F_{\lambda 2}$ representing a structure factor of the same indices but at a wavelength where f" is at a maximum.*

Bijvoet differences are alone sufficient to solve a macromolecular structure by the SAD method. MAD phasing however, relies not only upon Bijvoet differences but also on what are termed *dispersive differences.*

Dispersive differences are differences in intensity of equivalent reflections between data sets recorded at different wavelengths. Usually, datasets are recorded at three wavelengths; a *peak wavelength* using an incident X-ray beam energy equivalent to the experimentally determined absorption edge and the f" component is a maximum; an *inflection point wavelength* where the f' component is reduced to a minimum and a *remote wavelength* where the f' component is near its normal value. Dependent upon whether a high or low remote energy dataset is taken, f" is considerable, or at a minimum, respectively. Strictly it is the peak, inflection point and remote *energies* that are sought with the incident X-ray wavelength modulated to achieve these energies.

In a similar manner to Bijvoet differences, dispersive differences can be measured by comparison of equivalent reflections of a peak wavelength dataset with an inflection point dataset revealing the f' and f" contributions to each structure factor. Third or fourth datasets at remote wavelengths provides additional data upon which dispersive differences can be measured. Appropriate X-ray beam energies to allow selection of suitable wavelengths at which to collect peak, inflection point, and remote datasets can be determined by completing a fluorescence scan of the candidate crystal prior to data collection. Though theoretical values exist for f' and f" values for each heavy atom used in anomalous dispersion experiments they are only approximate since both parameters vary dependent upon the effects of neighbouring atoms in the protein structure. Fortunately, incident X-rays of energy capable of causing anomalous dispersion do not only cause excitation of a heavy atom with re-emission of an anomalously dispersed X-ray but also cause emission of lower energy electromagnetic radiation as fluorescence. This is convenient for the crystallographer as fluorescence is easily measurable and commensurate with X-ray absorption thereby allowing f' and f" to be plotted with incident X-ray energy. The actual wavelength used to achieve the experimentally determined energies is dependent upon the individual crystal properties and the beam-line setup and is therefore calculated specifically for each anomalous dispersion experiment (105, 106, 107).

2.4.4.5 Phasing using anomalous dispersion data

Central to phasing by SAD or MAD is calculation of the contribution to scattering of the heavy atoms only. The relatively complex diffraction signature of the protein can be reduced to a far simpler diffraction signature of a few heavy atoms in the unit cell. Currently locating more than fifty heavy atoms in the unit cell presents a problem due to the high complexity of the diffraction signature, though this is subject to change as phasing programmes improve.

If we return to the Patterson function described earlier it becomes clear that if we can measure X-ray scattering by heavy atoms only, the resultant Patterson map would consist of only a few peaks as the number of inter-atomic vectors would be relatively tiny corresponding to vectors between the heavy atoms only. For example, it will later be shown that the structure of the SV3CP, the subject of this thesis, contained 3125 atoms, producing a theoretical 9,762,500 peaks on the Patterson map. Ten of the 3125 atoms are the heavy atom selenium. By SAD and MAD phasing techniques it is possible to determine the contribution to X-ray scattering of only these ten seleniums and produce a Patterson map consisting only of inter-atomic vectors between the seleniums. The resultant difference Patterson map would contain only ninety peaks, well within the limits, allowing anomalous dispersion techniques to be applied in solving this simple structure of just ten atoms. By trial and error, the inter-atomic vectors described in the difference Patterson will reveal spatial locations for each of the heavy atoms in the unit cell. By suggesting an arrangement of atoms and calculating a Patterson map based upon these co-ordinates, a calculated Patterson map can be constructed that can then be compared with the experimentally obtained Patterson. If they agree a correct solution has been determined. Due to the still large number of permutations of atomic arrangements, even in this case just ten atoms, this is a computationally based step (105, 106, 107).

With the heavy atom locations determined, phases can be extended to protein atoms by firstly satisfying two equations to determine $F_{\lambda 1}^{+}$; a single structure factor of a heavy atom containing protein at a wavelength where f" is zero (*eq.7*).

$$F_{\lambda 1}^{+} = F_{\lambda 2}^{+} - F_{f'}^{+} - F_{f''}^{+} \qquad (eq.7)$$

Where $F_{\lambda 2}{}^{+}$ is a structure factor with the same indices as $F_{\lambda 1}^{+}$ but at a wavelength where f" is maximum; and $F_{f'}^{+}$ and $F_{f''}^{+}$ are the real and imaginary components of anomalous scattering respectively.

Equation 7 can be solved to offer two possible phase angles for $F_{\lambda 1}^{+}$ (the amplitude has of course been determined experimentally). The real and imaginary components of anomalous scattering $F_{f'}^{+}$ and $F_{f''}^{+}$ that modify $F_{\lambda 1}^{+}$ to become $F_{\lambda 2}^{+}$ when data is collected at wavelength where f" is maximum are always orthogonal (due to the phase lag of 90 ° following X-ray absorption). They are also constant in magnitude for a given element irrespective of environment so can therefore be looked up and are *not* measured from the experimental data. The phase angles of $F_{f'}^{+}$ and $F_{f''}^{+}$ can be calculated from their position in the unit cell; by this stage already known. To satisfy equation 1 partially we do not actually need to know the phase of $F_{\lambda 2}^{+}$ just yet, simply its magnitude and that has been experimentally determined. At this stage we have two possible phase angles for $F_{\lambda 1}^{+}$ reflected in the vectors F_a and F_b in (*Figure 2.15*). We must then use equation 2 to resolve this phase ambiguity by considering the Friedel partner of $F_{\lambda 2}^{+}$.

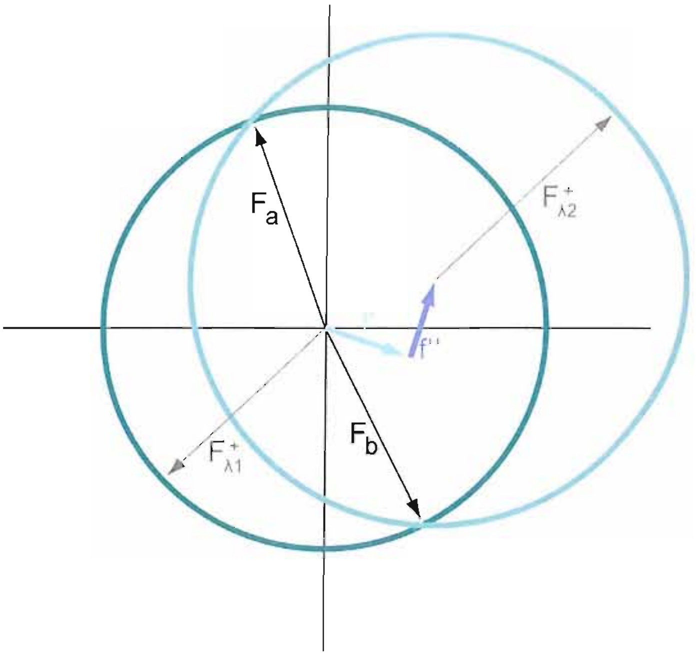


Figure 2.15: Geometric explanation of how equation 7 is satisfied. *Phases of $F_{f'}^+$ and $F_{f''}^+$ are derived from the positions of heavy atoms in the unit cell; in addition they are always orthogonal to each other. Magnitudes of $F_{f'}^+$ and $F_{f''}^+$ are constant for an element regardless of environment. If $F_{f''}^+$ is placed at the tail of $F_{f'}^+$ and a circle drawn of radius corresponding to the magnitude of $F_{\lambda 2}^+$ (obtained experimentally) and a second circle drawn at the origin of $F_{f'}^+$ of radius corresponding to the magnitude of $F_{\lambda 1}^+$, where the two circles intersect vectors F_a and F_b can be drawn; each corresponding to a possible phase solution for $F_{\lambda 1}^+$. It is clear to see that the phase of $F_{\lambda 2}^+$ is not required to draw the first circle in the correct position.*

A second equation (*eq.8*) must also be satisfied.

$$F_{\lambda 1}^+ = [F_{\lambda 2}^-] - F_{f'}^+ - (-F_{f''}^+) \qquad (eq.8)$$

Where $F_{\lambda 2}^-$ is the Friedel partner of $F_{\lambda 2}^+$.

If $F_{\lambda 2}^-$ is reflected across its axis its component vectors $F_{\lambda 1}^-$ and $F_{f'}^-$ and $F_{f''}^-$ must also be reflected and in doing so $F_{\lambda 1}^-$ becomes equal to $F_{\lambda 1}^+$ and $F_{f'}^-$ and $F_{f''}^-$ equal to $F_{f'}^+$ and $F_{f''}^+$ respectively. Equation *8* is simply equation *7* but with the substitutions described made. In an almost identical manner as equation *7* was solved (*Figure 2.15*) equation *8* can be solved to reveal two vectors of equal magnitude but different phase describing $F_{\lambda 1}^-$..

However the phase ambiguity is removed on this occasion since one of these vectors will have a phase very close to one of the two phases suggested by equation 7; and it is this vector that carries the correct phase.

So finally, to solve phases for protein structure factors; $F_{P,}$ we employ equation *9*, where F_H represents structure factors for heavy atoms only and are calculated after heavy atom sites are located by Patterson methods previously described (105, 106, 107):

$$F_P = F_{\lambda 1}{}^{+} - F_H \qquad (eq.9)$$

2.4.4.6 Using SOLVE/RESOLVE to solve SAD/MAD structures

The SOLVE/RESOLVE (118, 119, 120, 121, 122, 123, 142) programme suite allows automatic location of heavy atom sites and calculation of protein phases to provide phased electron density maps. The basic principle of its operation, Patterson maps to locate heavy atoms, are as those already described. The criteria SOLVE/RESOLVE uses to score possible solutions are of interest and will now be discussed. Each solution (a series of possible heavy atom sites) suggested by SOLVE is scored according to four criterion: *analysis of difference Pattersons*, *'free' self-difference Fourier analyses*, *non-randomness test* on the native Fourier and a *figure of merit of phasing*, to arrive at a final score that is reflective of the viability of a suggested solution. This overall score is termed the *Z-score*.

Analysis by difference Pattersons involves comparison of calculated Patterson maps for a set of possible heavy atom sites with the actual Patterson map experimentally obtained; the *rms* deviation of each peak height is then calculated.

Difference Fouriers can be calculated once initial heavy atom sites have been suggested. SOLVE utilises an adaptation of difference Fouriers to score a solution by employing a technique termed “free” self difference Fourier transformation. From all heavy atom locations, except the site under investigation, protein phases are derived and the subsequent Fourier calculated. This protein density is then used to calculate a difference Fourier comprising only of terms describing electron density of the original heavy atoms used in phasing plus the heavy atom site under investigation. The Fourier is then inspected and the peak height of the omitted heavy atom is recorded. It follows that if the site under investigation is suspect, this method removes it from use in phasing and hence, incorrectly, influencing the Fourier. If a site is completely incorrect, density should be absent at its location in the difference Fourier. This process is repeated for all heavy atom sites of a solution.

The non-randomness test on the native Fourier i.e. the complete Fourier describing all contents of the unit cell, acts to determine whether electron density is randomly distributed within the unit cell. Solvent regions have a low level of density, therefore the standard deviation of density within these regions is also low. Protein regions have moderate to high levels of electron density, therefore the standard deviation of electron density within these regions is comparatively high. SOLVE samples the electron density throughout the unit cell to determine whether the standard deviation is low, indicative of continuous density and a poorly phased map, or high, indicative of discontinuous density and a well phased map. As a further measure, SOLVE samples areas of adjacent density to determine whether the density is contiguous or not; along with the measure of non-randomness, this parameter is expressed as a single term; the ‘correlation coefficient’.

As a fourth and final criterion, SOLVE utilises a phasing figure of merit measure that if high and commensurate with favourable scores for the other three criteria will likely be reflective of a good solution.

As a rule of thumb, a Z-score of above 20 and a phasing figure of merit above 0.5 are indicative of a good solution. Following heavy atom site refinement and associated phase improvement in RESOLVE, a phasing figure of merit of 0.65 at minimum should be expected.

2.5 Cryo-crystallography

A relatively recent advance in protein X-ray crystallography has been the use of cryo-cooling, where protein crystals are frozen instantly upon submersion in liquid nitrogen prior to data collection conducted at approximately 100 K (103). Cryo-cooling has a number of advantages over data collection without cooling, the most valuable being the reduction of radiation damage the crystal suffers during data collection (135, 136). The resonance induced in protein molecules and subsequent generation of free radicals when exposed to the high intensity of an X-ray beam can result in collapse of the lattice of a protein crystal. The resultant disorder within the crystal reduces its ability to diffract X-rays. Radiation damage is manifest in a number of ways; loss of diffraction lowers data resolution and may result in incomplete data, or, specific structural damage such as de-carboxylation of glutamate and aspartate residues, or breakage of disulphide bridges (90). At cryogenic temperatures, radiation damage is greatly reduced. In addition cryo-cooling allows protein crystals to be harvested and then stored safely when in peak condition; crystals can often form then degrade if left at non-cryogenic temperatures. Further, the mosaicity, a measure of disorder, of a crystal is reduced at cryogenic temperatures compared to room temperature (104).

In theory, flash freezing at cryogenic temperatures should form a glass of the water of the crystallisation buffer surrounding the crystal i.e. prevent the water forming a regular lattice as displayed by ice. In practice, this does not happen to the extent wished and therefore a cryo-protectant is added to the buffer containing the crystal. The most common cryo-protectants used include glycerol, ethylene glycol, MPD and PEG; the cryo-protectant acts as an anti-freeze and so goes some way to preventing ice crystals forming upon freezing of the protein crystal (*Figure 2.16*).

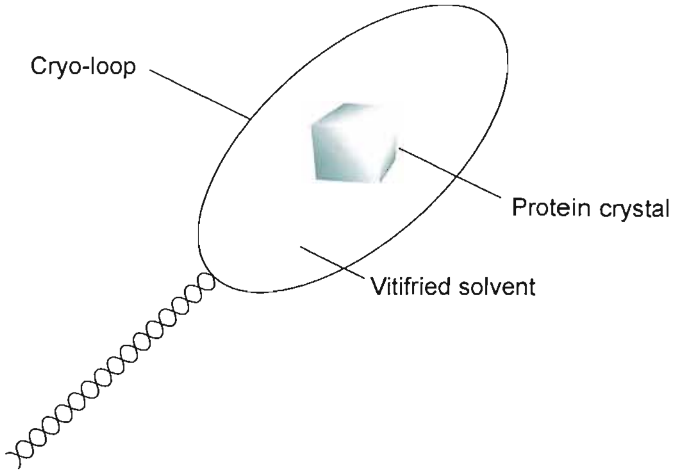


Figure 2.16. A crystal mounted in a cryoloop. *A crystal is removed from the drop in which it formed by means of a cryo-loop then transferred to a drop of the same solvent. Gradually, a suitable cryoprotectant, often glycerol at approximately 30% final concentration, is mixed into the drop. The crystal is then removed from the smaller drop using the cryo-loop. The crystal becomes suspended in the cryo-loop by surface tension of the surrounding crystallisation buffer. The cryo-loop is then submersed in either liquid nitrogen immediately or liquid ethane then liquid nitrogen, if a two step freezing is adopted. A clear frozen drop containing the crystal is indicative of effective freezing and solvent vitrification. If ice forms, the drop will have a cloudy appearance. Ice may lead to significant X-ray diffraction manifest in 'ice rings' (concentric circles at approx 5 Å) on diffraction images that can undesirably obscure diffraction due to protein. Note that disulphide bridge breakage due to radiation damage may be reduced by addition of 0.5M ascorbate to the cryo-protectant solution (90).*

2.6 Collection of X-ray data

In its simplest terms data collection involves placing a suitable protein crystal in the path of a narrow X-ray beam and recording the resultant diffraction. The essential instruments required are therefore; an *X-ray source*, a *goniometer* (a means of orientating and rotating the crystal in the X-ray beam) and an *X-ray detector* (*Figure 2.17*). X-Ray sources maybe small rotating anode sources or one of the approximately 70 worldwide high energy synchrotron sources that provide exceptionally intense and tightly focussed

X-ray beams. Various X-ray detectors have been used in the past ranging from relatively low tech X-ray film through multi-wire detectors and image plates (still seen on small rotating anode set-ups), though now most crystallography facilities use charge coupled device detectors (CCD) a modern high resolution type of detector.

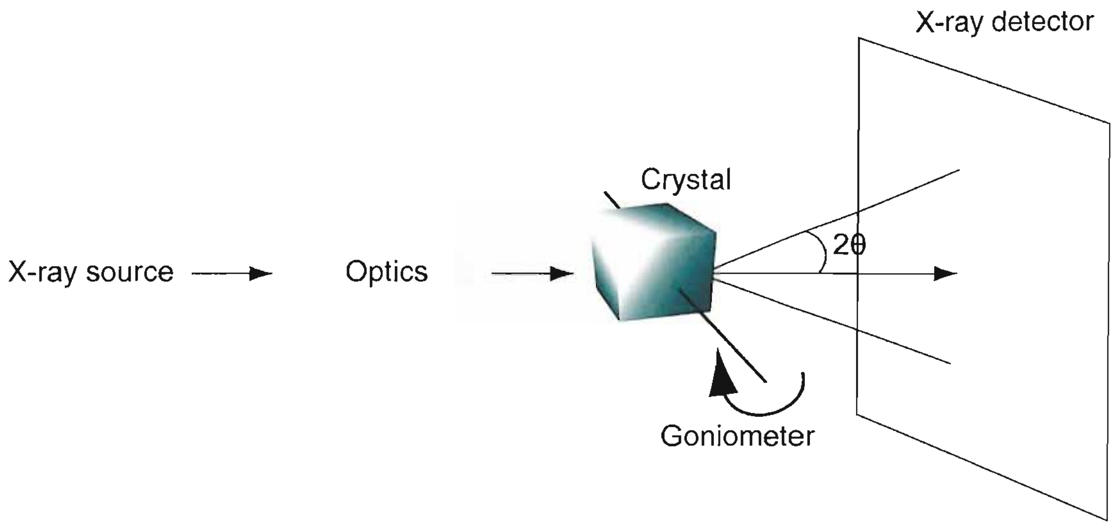


***Figure 2.17: A diagrammatic representation of the basic experimental set up for X-ray diffraction data collection**. Note that the crystal is kept at cryogenic temperatures by being placed in a stream of liquid nitrogen know as the cryo-stream. The majority of X-rays comprising the X-ray beam actually remain un-diffracted and pass straight through the crystal hence the need for a 'backstop'; a small metallic block capable of absorbing X-ray energy and preventing damage to the sensitive detector.*

Over the past ten years, third generation synchrotron sources have emerged that have been designed specifically with protein X-ray crystallography in mind these include: the ESRF in Grenoble, France; the SLS at the Paul Scherrer Institute, Switzerland; SPRING8, Japan; and the partially functional DIAMOND light source, Oxford, UK. Such sources provide a highly parallel, mono-chromated beam i.e. possess a very small wavelength spread and, in some cases, the X-ray beam is wavelength tuneable (125). In comparison to rotating anode sources, a synchrotron X-ray beam has an amazingly highly brilliant and focused beam of no more than 0.1 mm^2 in width. The typical brightness of synchrotron radiation is 10^{20} photons sec^{-1} mm^{-2}, compared with the typical output of 10^6 photons sec^{-1} mm^{-2} of a rotating anode source (113). These properties provide an X-ray source that is ideal for collecting high resolution X-ray diffraction data sets, as the higher strength the X-ray beam the stronger the diffraction from a protein crystal will be (to a point) (138).

A synchrotron source exploits the phenomenon of the production of X-ray radiation by accelerating electrons that are already travelling at relativistic speeds (112). Acceleration is achieved by forcing the path of such electrons to travel in a circular motion maintained by bending electromagnets, producing X-ray radiation of moderate strength over a range of wavelengths though usually the optics of a beam-line set-up fed by a bending magnet will filter the X-ray beam to a single wavelength. For protein crystallography work, this will be a wavelength of approximately 0.97 Å. Note that the "optics" equipment of an X-ray beam-line set-up include; attenuators, monochromators, mirrors and collimators to produce a coherent X-ray beam of known wavelength (114, 129).

Additional deviation to produce higher intensity radiation over a range of wavelengths is provided by the use of insertion devices. Insertion devices are also electromagnets and of two types: wigglers and undulators. Wigglers cause the greatest deviation in direction of travel of the electrons and generate X-rays of multiple wavelengths. X-Ray radiation produced by wigglers is comprised of wavelengths between 0.3 Å and 3.5 Å that may be filtered to restrict wavelength and to modify the X-ray beam energy. Specialized "optics" allow the precise selection by the user of X-ray wavelength making beam-lines fed by wigglers suitable for use in experimental phasing experiments where successive datasets need to be recorded at a range of X-ray energies. Undulators cause a lesser deviation producing X-rays of very similar wavelengths resulting in a very intense beam suitable for fixed wavelength experiments (114).

For sufficient diffraction data to be collected to allow a structural solution, it is not sufficient to expose the protein crystal to an X-ray beam whilst in a single orientation only. Instead the crystal must be rotated through a small angle on an axis perpendicular to the X-ray beam with diffraction data comprising a single diffraction image usually collected over a 1° rotation (smaller rotation angles can be used to minimise large overlaps of reflections if seen). Each diffraction image comprises of reflections as a result of diffraction throughout the 1° of rotation.

Raw diffraction data comprises of a set of individual images separated by the chosen rotation angle; typically this will be 180 diffraction images each separated by 1° and therefore covering 180° of rotation of the crystal in the X-ray beam. Though diffraction data spanning a larger degree of rotation, sometimes spanning multiple crystals, maybe collected for experiments where phase information is sought i.e. SAD/MAD experiments where a high degree of data redundancy is desirable (105, 106, 107).

2.7 Overview of the computational approach to structural solution

2.7.1 Data processing

Data processing is the computational step taken to reduce the raw diffraction data to a list of reflection spot intensities with associated errors as a set of phase-less structure factor amplitudes. When the structure factor amplitudes are coupled with associated phases the data represents a Fourier transform of the protein electron density.

Data processing can be divided into three main stages

1. The determination of unit cell parameters. These include unit cell dimensions; cell edge lengths a, b, c and cell angles α, β, and γ, identification of space group, estimation of mosaicity and the crystal orientation. The unit parameters are essential to allow reflection positions to be predicted and subsequently to measure and record the intensity of each reflection. This process is termed 'indexing'.
2. The integration of diffraction images – necessary so that all reflections and intensities, including partially recorded reflections (recorded over multiple images), can be recorded.
3. Data reduction where all reflections are placed upon a common scale, reflections that have been recorded multiple times are merged (the extent depends upon crystallographic symmetry of the crystal data though during SAD/MAD

experiments reflections may remain unmerged), with outlying reflections and overloads (reflections that have exceeded the dynamic range of the detector) rejected.

MOSFLM (116, 117) is a widely used programme that with a low level of user input can complete steps 1 and 2 of data processing. MOSFLM utilises a positional residual ($R_{ms)}$ that indicates the accuracy of the determined unit cell parameters. For a reasonable estimate of the cell parameters the positional residual should be below 0.15 mm for a CCD type detector or below 0.25 mm for image plate type detectors. The positional residual is the standard deviation of the actual spot position on the diffraction image from the predicted spot position as defined by the cell parameters. Correspondingly, the positional residual will be higher for weak data as reflection positions are not as well defined.

SCALA, part of the CCP4 programme suite (115), will perform data reduction upon the MOSFLM output to produce a list of *h, k, l* values with the associated scaled and merged intensities and individual error in intensity. It is essential to scale reflection intensities to take account of crystal degradation as a result of X-ray damage, beam intensity fluctuation during data collection and reflection intensity variation during rotation of the crystal; a thicker part of the crystal will possess the ability to more strongly diffract X-rays. Further, SCALA employs *Lorentz correction*; correcting for the differing times taken by the reciprocal lattice points to traverse the Ewald sphere. At this stage a valuable measure of data quality is the R_{merge} statistic produced by SCALA. The R_{merge} value is a measure of the relative intensities of symmetry related reflections. R_{merge} is a recent improvement over the statistical measure of data quality: R_{sym}, since R_{merge} takes into account multiplicity within the dataset i.e. if the crystal possesses a high degree of symmetry it will generate a larger number of redundant reflections that would artificially raise the R_{sym} value. An R_{merge} value of below 10% is indicative of good quality data. Further, an R_{merge} of 10% can be broadly interpreted as symmetry related reflection intensities only varying by 10%.

2.8 Model building and structural refinement

2.8.1 Building the initial model

Whichever the phasing technique used initial phases are just estimates, though the initial electron density maps obtained from these phase estimates will be of sufficient quality to allow the majority of a protein structure to be built into the density with relative confidence. In the case of structures solved by MR, the structural homology between model and target structures would have to be sufficient to phase and, as such, provides a large amount of information when building the target structure into density. Experimentally phased data (SIR, MIR, SAD, MAD, SIRAS, MIRAS experiments) requires the protein model to built into density from scratch, although programmes do exist (RESOLVE (142), MAID (145), ARP/wARP (132)) to automate this process they do not always provide a decent model, usually resulting in a partial model with the remainder required to be built by hand.

From an initial model, structure factors can be calculated; values of $F_{(calc)}$. $F_{(calc)}$'s are the reverse Fourier of the model structure. Refinement involves improving the agreement of these calculated structure factors with the actual structure factors observed experimentally $F_{(obs)}$. The overall aim of refinement is to minimise the term described by *eq.10*.

$$\Delta F = \left| F_{(obs)} - F_{(calc)} \right| \qquad \textit{(eq.10)}$$

Since the values for $F_{(calc)}$ are calculated, based upon the current model structure, with inherent inaccuracies the agreement with $F_{(obs)}$ can be improved by refinement of model atomic positions, temperature factors and occupancies. Refinement involves iterative rounds of manual adjustment of atomic positions in real space followed by automated computational refinement comprising of further positional, temperature factor (grouped, individual, and anisotropic; dependent upon data resolution) and occupancy refinement. At the time of writing, programmes commonly used for automated refinement include; CNS (124), REFMAC 5 (130), SHELX (131) and ARP/wARP (132).

Central to all refinement programmes is the procedure of least squares refinement (141) where the squares of differences between values for $F_{(calc)}$ and $F_{(obs)}$ are minimised and can be expressed by *eq.11*.

$$\Phi = \Sigma W_{hkl} (| F_{(obs)} | - | F_{(calc)} |)^2_{hkl} \qquad (eq.11)$$

Where Φ is the sum of the squared differences between $F_{(obs)}$ and $F_{(calc)}$, and W_{hkl} is a weighting factor reflective of the reliability of a given structure factor.

When dealing with diffraction data of 2.0 Å resolution or lower, least squares refinement must be completed with stereo chemical restraints applied. That is, a library of stereo chemical data is applied by the refinement programme to ensure all bond lengths and angles are maintained within established limits. In addition to bonds lengths and angles further restraints including; the planarity of ring side chains; torsion angles; temperature factor; and close contact restraints for non bonded atoms are also applied. The effect of utilising such restraints during refinement is to overcome the poor observations to parameter ratio experienced in solving all but high resolution structures. Since restraints offer information by placing the parameters described within set limits they effectively act as additional observations. A high observation-to-parameter ratio is essential for refinement by the least squares method.

A shortcoming of least squares refinement is that is can only proceed in an energetically decreasing direction. That is, from a relatively high energy start point least squares refinement will seek the closest minimum energy state and this may or may not be an appropriate direction towards a global energetic minimum. If this occurs, refinement is said to have become stuck in a local minimum. This is where manual model building becomes essential to redirect refinement. The relatively large changes that can be manually made to a structure can lift the structure out of a local minimum, over come an energy barrier and set refinement back on course to converge to a global energetic minimum and a correct structure. Hence structural solution involves this iterative process of computational refinement followed by manual 'refinement', or 'model building' as it is more commonly termed.

To aid the crystallographer in the model building process two types of electron density map are usually calculated; the $2F_o$-F_c map and the F_o-F_c difference map. As the name $2F_o$-F_c implies the $2F_o$-F_c map is 2 times the observed minus the calculated electron density and represents the density of the current model defining the main chain and side chain atoms. The F_o-F_c map displays the difference between the observed and calculated density and as such aids in identification of density not explained by the model. Positive density in the F_o-F_c map is indicative of observed density not accounted for by the model; negative density is indicative of the model being misplaced in a region where density does not exist in the observed data.

The success of refinement can be tracked using two statistical measures; the traditional crystallographic R-factor and the R_{free} value (99). Both measures are reflective of the level of agreement between $F_{(obs)}$ and $F_{(calc)}$ and therefore the accuracy of the structural model. The R-factor is described by *eq.12*.

$$R = \Sigma \left| F_{(obs)} - F_{(calc)} \right| / \Sigma F_{(obs)} \qquad (eq.12)$$

Following the first round of refinement with a preliminary structure, an R-factor of approximately 0.4 would be expected. For a good quality structure the final R-factor would be expected to equal approximately 0.2; though values as low as 0.15 are achievable and R-factors as high as 0.25 may be acceptable depending upon the impact of the structure! It should be noted that if values for $F_{(calc)}$ were randomly calculated, an R-factor of 0.6 would result.

Alongside the R-factor, an R_{free} value is often quoted. The R_{free} value is so-named as a portion, usually 5-10%, of reflections that are removed from the dataset at the beginning of the refinement process. The R_{free} is then calculated in a similar manner to the R-factor however only upon this reserved portion (the “test set”) of reflections. The principle of R_{free} being how well the refined model predicts the test set reflections and therefore removes the probability of model bias during refinement as long as it used in conjunction with the traditional R-factor as indication of successful refinement. The R_{free} value would be expected to be between 2-5% higher than the R-factor for a decent structural solution (105, 106, 107).

3.0 Synthesis of a series of chromogenic peptides and a Michael acceptor peptidyl inhibitor of SV3CP

3.1 An introduction to solid phase peptide synthesis

The use of a solid support onto which amino acid residues could be coupled for the organic synthesis of peptides was pioneered by R.B. Merrifield during the early 1960's (148). Merrifield proposed a method that circumvented the time consuming issues associated with solution phase synthesis of peptides entailing the isolation and purification of peptide species following the coupling of *every* residue to the peptide chain.

Merrifield's solid phase synthesis of peptides involves the attachment of the *C*-terminal residue of a peptide sequence to an insoluble polystyrene resin (the solid phase) *via* the formation of an acid labile peptide-benzyl ester linkage. With the amine group of the residue coupled to the resin protected by a tertiary butoxycarbonyl (Boc), coupling of further residues and premature extension of the peptide chain is prevented until dersired. Following the initial coupling of the *C*-terminal residue to the resin the acid labile Boc protecting group can be cleaved by addition of trifluoroacetic acid (TFA) to allow coupling of the next *N*-terminal Boc protected residue in the peptide sequence (*Figure 3.1*). Thereby Merrifield's solid phase peptide synthesis (SPPS) method allows the controlled extension of the peptide chain in the C to *N*-terminal direction by a single residue at a time. Further, the attachment of the peptide chain to an insoluble bulky resin allows easy removal of excess reagents and retention of the peptide following each reaction stage. Following synthesis of the desired peptide, it may be cleaved from the resin by the addition of the strong acid, hydrogen fluoride (HF) (*Figure 3.1*). This technique adopts a strategy of graduated acidolysis, i.e. the *N*-terminal Boc protecting group possesses a higher acid lability than the peptide benzyl ester bond linking the *C*-terminal residue to the resin. Under relatively weakly acidic conditions, the *N*-terminal protecting group can be cleaved and the peptide extended then, under strongly acidic conditions, the completed peptide can be cleaved from the resin. To prevent the formation of side chain extended derivatives the amino acid functional side groups are usually protected by benzyl based protecting groups that possess relatively weak acid lability and are therefore only removed during the final reaction step; cleavage of the peptide-resin linkage under strongly acidic conditions (*Figure 3.1*) (151).

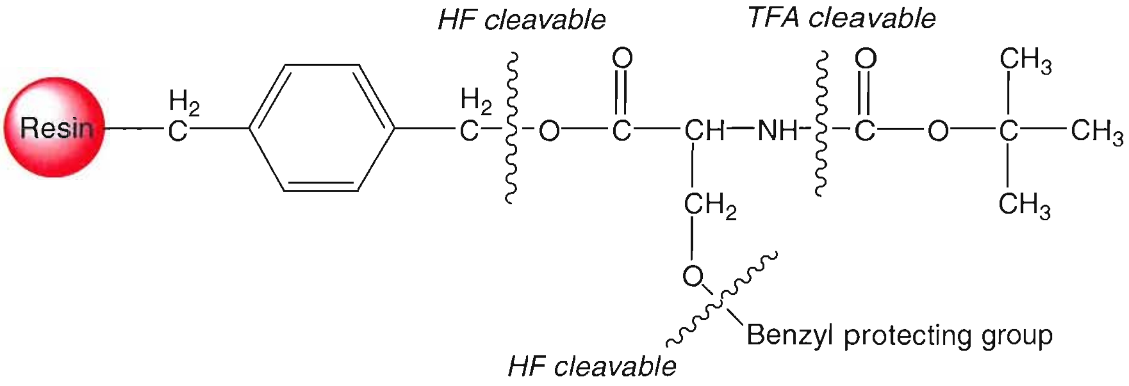


Figure 3.1: Overview of reaction conditions required for cleavage of protecting group and peptide-resin linkage using the Merrifield SPPS method.

Despite being an effective technique for peptide synthesis, the Merrifield method has widely been replaced with the milder reaction conditions provided by Fmoc SPPS: a technique largely based upon Merrifield's original method but employing a base labile 9-fluorenmethoxycarbonyl (Fmoc) *N*-terminal protecting group (149, 150, 151). The basic conditions required for Fmoc cleavage allow the approach of graduated acidolysis, as employed using Boc protected amino acids, to be dispensed with and therefore acidic conditions are required only for the cleavage of the peptide-resin linkage and side chain protecting groups (e.g. tertiary butyl, trityl, methyl: many exist) (*Figure 3.2*). In doing so, the inconvenience and danger associated with working with HF is removed. Most importantly, the milder reaction conditions allow the convenient synthesis of more difficult peptides not possible using the Merrifield method where Boc protected amino acids are used. Long peptides that would gradually be cleaved from the resin prior to completion due to the repeated use of acidic conditions in cleaving Boc protecting groups can be synthesised using the Fmoc approach. Similarly Fmoc SPPS lends itself to the synthesis of bespoke peptides that contain groups prone to cleavage under repeated acidic conditions e.g. Arginine.

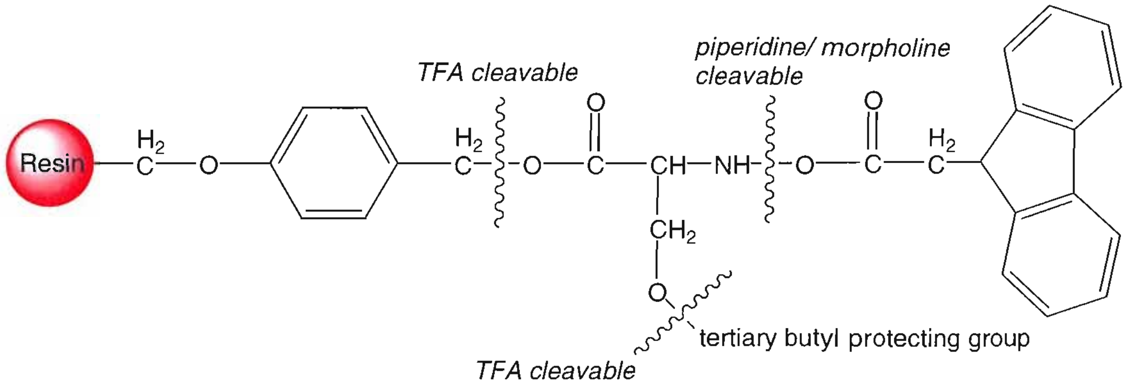


Figure 3.2: Overview of reaction conditions required for cleavage of protecting group and peptide-resin linkage using the Fmoc SPPS method.

In the same manner as the original Merrifield method, SPPS by Fmoc chemistry involves the step by step addition of *N*-terminal Fmoc protected amino acids to a peptide of desired sequence whilst the *C*-terminal of the growing peptide remains attached to a support resin. Extension of the peptide chain becomes the repetition of a cycle of reactions; cleavage of the *N*-terminal Fmoc protecting group of the residue attached to the resin, followed by the addition of the next Fmoc protected residue through activation of it's *C*-terminal carboxyl group and simultaneous coupling. The bases, *piperidine* and *morpholine* are commonly used for cleavage of the Fmoc protecting group with *trifluoroacetic acid* (*TFA*) used in the final peptide-resin cleavage; the reagents *HOBt* and *TBTU* are used to favour coupling during chain extension by activation of the incoming carboxyl group of the subsequent residues (*Figure 3.3*) (151).

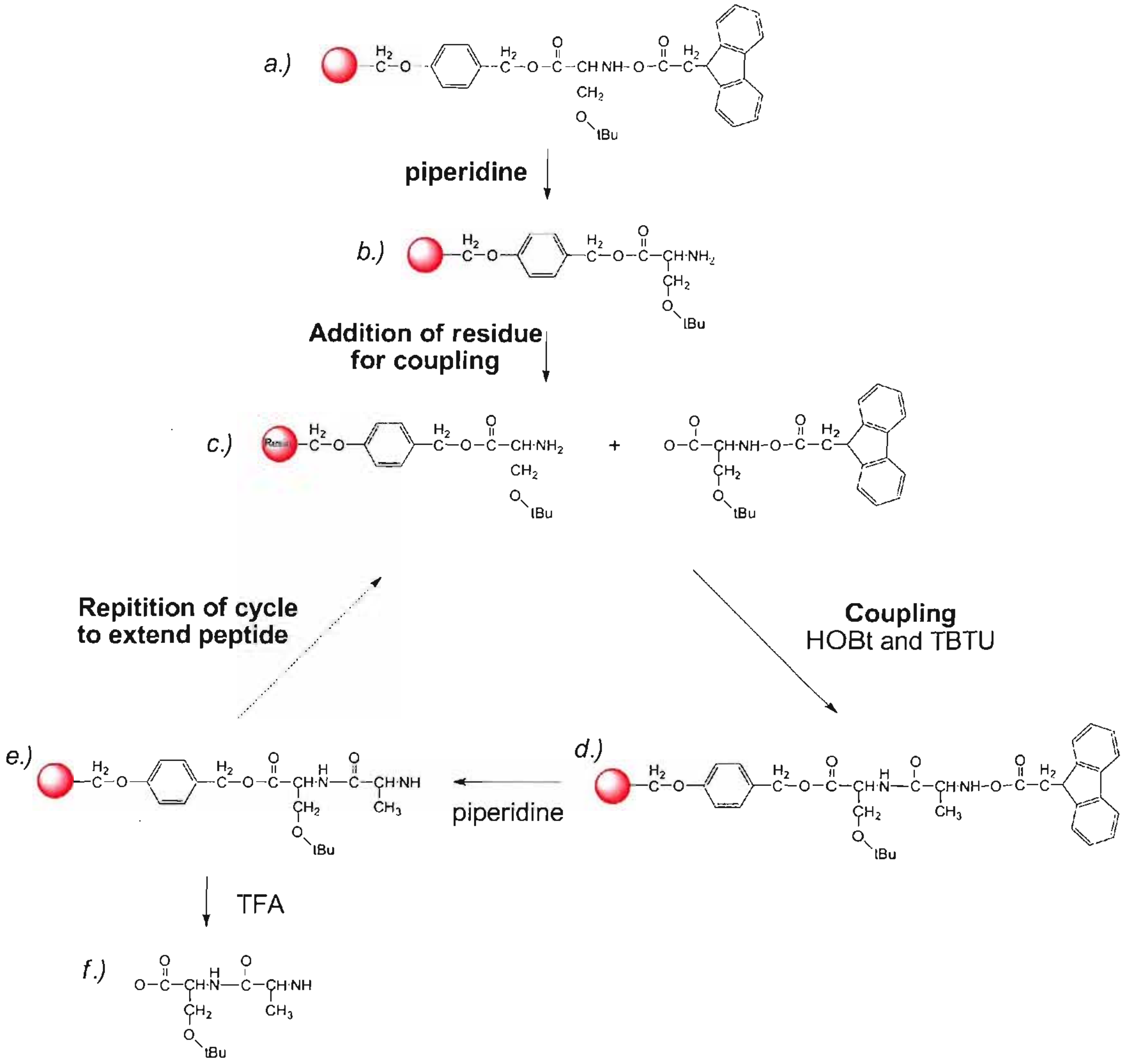

a.)
piperidine
b.)
Addition of residue for coupling
c.)
+
Repitition of cycle to extend peptide
Coupling
HOBt and TBTU
e.)
piperidine
d.)
TFA
f.)
tBu

***Figure 3.3: Overview of Fmoc SPPS; extension of the peptide chain involves the repetition of a cycle of reactions**. a.) Representation of a single Fmoc protected residue coupled to a solid support resin. b.) Cleavage of the N-terminal Fmoc protecting group under basic conditions by use of piperidine. c.) and d.) addition of a Fmoc protected residue to the deprotected C-terminal residue, and coupling of residues in presence of coupling agents HOBt and TBTU. e.) The Fmoc protecting group of the N-terminal residue is cleaved under basic conditions. Chain extension is achieved by successive rounds of deprotection and coupling, until f.) The resin-peptide linkage and side chain protecting groups are cleaved from the peptide using TFA to yield the final product* (151).

3.2 Synthesis of a range of chromogenic peptides for assay of SV3CP activity

The integrity of most biological systems is maintained by factors effecting interaction of enzymes and substrate belonging to that system. Amongst these factors are: the spatial and temporal localisation of enzyme and substrate; their absolute and relative concentrations and requirement of enzyme co-factors. Most dramatically, however, it is the specificity for substrate at the enzyme active site that affects enzyme activity and substrate turnover. Characterisation of enzyme substrate specificity can reveal the fundamentals of enzyme substrate interaction, detail of enzyme function and allow for dissection of the biological system to which the enzyme belongs. Further, probing of substrate specificity may also reveal information that may be utilised during the design of synthetic substrates of use during study of enzyme activity and also inhibitory compounds suitable for development as therapeutic agents.

A series of peptides were synthesised that provided a convenient colorimetric assay of the SV3CP activity. The chromogenic synthetic peptides: Ac-QLQ-*p*NA, Ac-FQLQ-*p*NA, Ac-EFQLQ-*p*NA and Ac-DEFQLQ-*p*NA were synthesised using a combination of Fmoc SPPS chemistry and synthetic chemistry techniques. Each peptide mimics a natural recognition sequence of the SV3CP in the 200kDa ORF1 polyprotein (see Section 1.4); the cleavage site proven in an *in vitro* expression system to experience the greatest rate of cleavage (58).

Each peptide has an acetylated *N*-terminus with a *C*-terminal linked *para*-nitroaniline group (152, 153). With the *C*-terminal glutamine occupying the protease active site P1 position and the *p*NA group occupying the P1' site; the peptide bond is located for hydrolysis by SV3CP. The uncleaved chromogenic peptide is colourless, however, upon cleavage (*Figure 3.4*) the *C*-terminal *para*-nitroaniline (*p*NA) group is protonated to yield the group *para*-anitroanilide. *Para*-nitroanilide is, in one state, strongly yellow in colour, the intensity of which can be measured spectrophotometrically at 405 nm. Free *para*-nitroaniline resonates between two states as electron density shifts around the molecule. One of these states (*Figure 3.4c*) displays the intense yellow colour described. The chromogenic peptides therefore present a means of quantitatively measuring SV3CP activity spectrophotometrically.

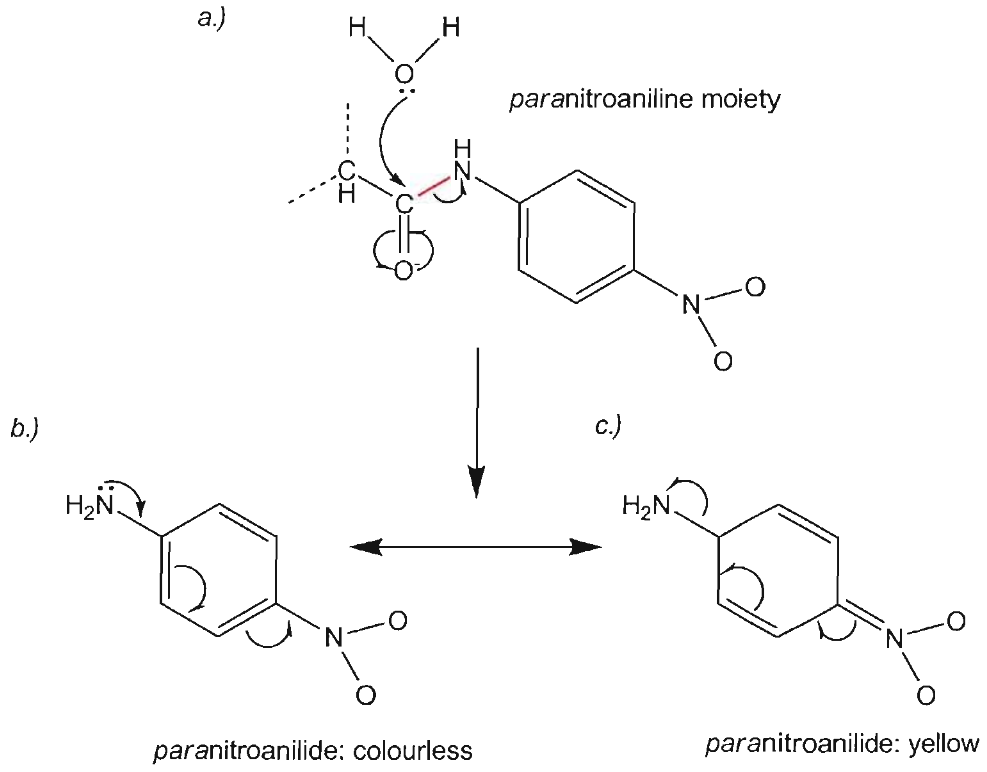

a.)
b.)
c.)
*para*nitroaniline moiety
*para*nitroanilide: colourless
*para*nitroanilide: yellow

Figure 3.4: Cleavage of the C-terminal pNA group yields an intensely yellow species. *a.) Hydrolysis of the pseudo-amide bond (coloured red) between the C-terminal glutamine and para-nitroaniline moiety yields para-nitroanilide. Para-nitroanilide resonates between the two states b.) and c.). The state shown in c.) is yellow in colour and its intensity measurable spectrophotometrically at 405 nm.*

3.2.1 Synthetic peptide synthesis

The peptide portions of the substrate derived synthetic peptides Ac-QLQ-*p*NA, Ac-FQLQ-*p*NA, Ac-EFQLQ-*p*NA and Ac-DEFQLQ-*p*NA were synthesised *via* established Fmoc solid phase peptide synthesis (SPPS) techniques (*Figure 3.5*).

Common to all synthetic peptides is the *C*-terminal chromogenic derivative glutamine, possessing the *C*-terminal linked chromophore: *p*NA. This presents an issue during peptide chain extension using SPPS techniques. Normally the *C*-terminal residue of a peptide to be synthesised by SPPS would be attached to the support resin *via* its *C*-

terminus. In this instance *p*NA must be positioned at the glutamine *C*-terminus to be positioned correctly in the SV3CP active site cleft for cleavage. As such the *C*-terminal residue must be linked *via* its side chain to the support resin. To solve this issue, the *p*NA derivative of *glutamate*, rather than *glutamine*, is synthesised, then linked to a Rink amide MBHA resin (154, 155) *via* its side chain carboxylic acid. Upon cleavage from the resin the side chain is aminated to yield a *glutamine* side chain. Specifics of synthesis follow.

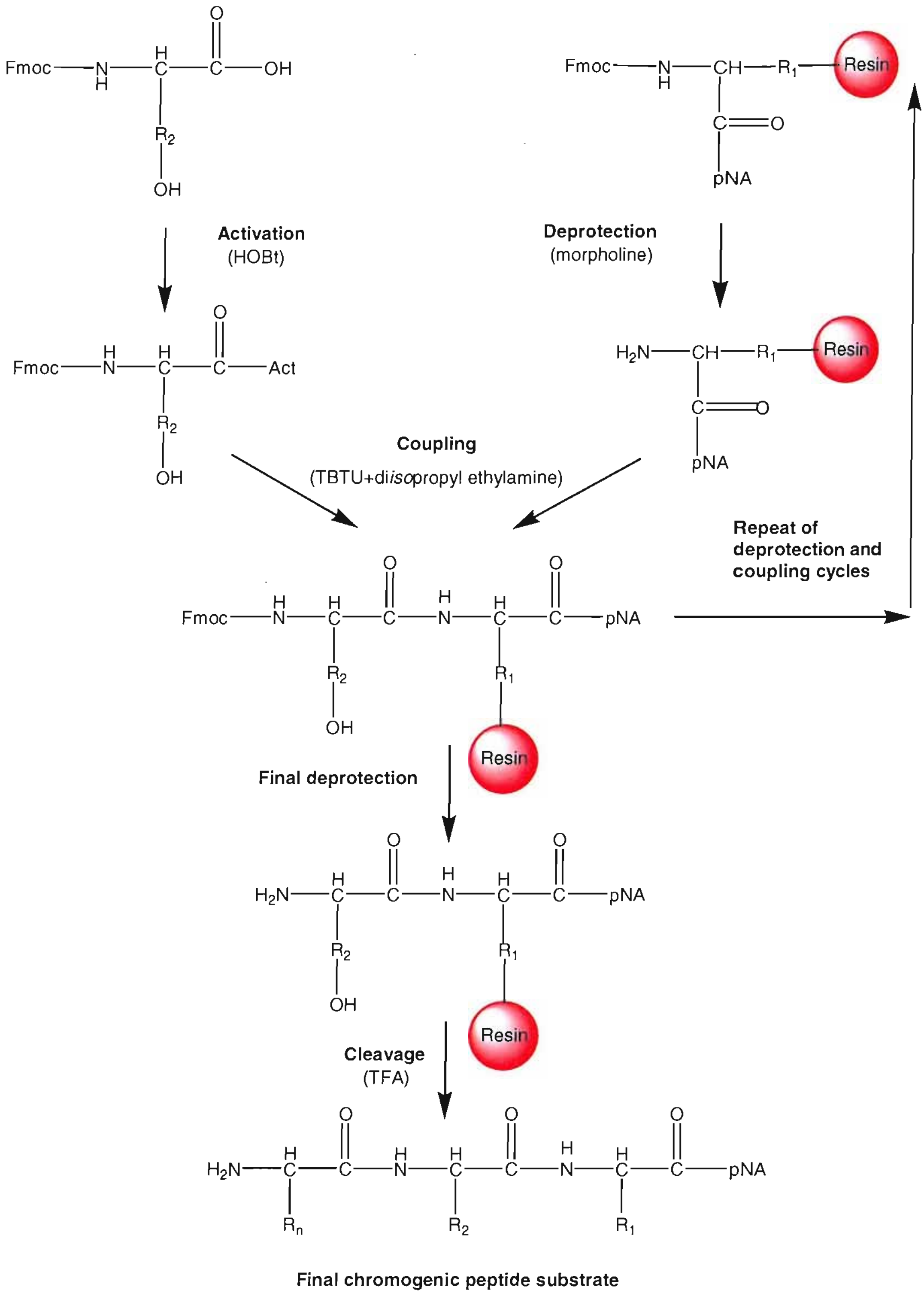


Figure 3.5: Overview of strategy for synthesis of the synthetic chromogenic peptides Ac-QLQ-pNA, Ac-FQLQ-pNA, Ac-EFQLQ-pNA and Ac-DEFQLQ-pNA.

3.3 Synthesis of Fmoc-glu(OH)-*p*NA

The first step in the preparation of all synthetic peptides was the synthesis of the *C*-terminal species: the *C*-terminal *p*NA derivative glutamate from *N*-terminal Fmoc protected, side chain tertiary-butyl (OtBu) protected glutamate (153, 156).

3.3.1 Protocol for production of Fmoc-glu(OtBu)OH-*p*NA

To 30 ml dry THF in a 250 ml round bottomed flask (RBF), the equivalent weight of 4 mmoles Fmoc-L-glu(OtBu)OH was added and slowly stirred to allow for complete dissolution. Following stirring, the activating agents were added; 1 ml of *diphenylphosphinic chloride* distilled under argon plus 750 µl of *N-ethylmorpholine*. The reaction mixture was then stirred slowly for 25 minutes to allow for *C*-terminal activation of the Fmoc-L-glu(OtBu)OH. The equivalent of 20 mmoles of 4-nitroaniline was then added to the reaction mixture and the mixture stirred for 6 hours to allow coupling of the *4-nitroaniline* to the *C*-terminus of the Fmoc-L-glu(OtBu)OH (*Figure 3.6*) (156). Following the reaction, the mixture adopted an opaque yellow appearance. It was now necessary to remove the solvent: THF, solvent evaporation was achieved using a rotary evaporator (RE) with warming at 40°C (156).

The original published papers (153) stipulate that the reaction mixture should be stirred on ice during coupling to prevent racemisation of product. In practice only 0.01% of glutamate will form the D-enantiomer therefore not significantly effecting yield of the L-enantiomer. Further, any product consisting of the D-enantiomer may still be separated later during purification by reverse phase chromatography.

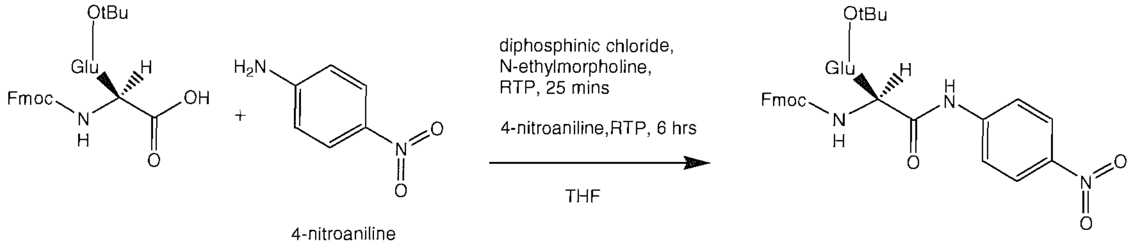


Figure 3.6: Reaction conditions for the C-terminal modification of N-terminal Fmoc protected and OtBu side chain protected glutamate.

3.3.2 The work-up

The product was washed to remove impurities; unreacted Fmoc-L-glu(OtBu)OH, 4-nitroaniline and some residual HCl. As Fmoc-L-glu(OtBu)*p*NA is an uncharged species it will dissolve in an organic solvent e.g. ethyl acetate, but remain insoluble in an inorganic polar solvent e.g. brine (a saturated solution of NaCl in water). When ethyl acetate and brine are added to the reaction mixture, two immiscible layers form. The upper layer, the organic phase, is ethyl acetate with the dissolved product Fmoc-L-glu(OtBu)OH-*p*NA; the lower layer, the inorganic phase, is brine with the unreacted and charged species; 4-nitroaniline and Fmoc-L-glu(OtBu)OH. Correspondingly, the resulting product following evaporation of THF, was dissolved in 15 ml of ethyl acetate to allow for complete dissolution. The dissolved product was then transferred to a separating vessel (of non-equilibrating type). A volume of a saturated solution of brine equivalent to half the volume of ethyl acetate was added and the separating vessel shaken vigorously to achieve an emulsion that was then allowed to settle and eventually separate. Two immiscible layers were observed: a translucent yellow upper layer (product in ethyl acetate) and a non-homogenous opaque bubbly grey lower layer (unreacted species in brine). The appearance of the lower layer was due to small amounts of product dissolved in ethyl acetate entering this aqueous phase. Leaving the contents of the separating vessel to stand for a short period allowed the majority of bubbles (bubbles of ethyl acetate covered with product) to burst and release the product back into the upper layer. The brine layer was then decanted to leave the organic phase: the product dissolved in ethyl acetate.

Unfortunately the separation of product from impurity was not as complete as desired. The unreacted species only possess a slight charge and therefore some dissolution of impurities in the organic phase occurs. Repeating the separation so that the reaction mixture had been washed in brine three times reduced the amount of impurities remaining in the organic phase.

The product was then dried over anhydrous sodium sulphate to remove any water present in the ethyl acetate (156).

3.3.3 Protocol for drying products over sodium sulphate

A sufficient amount of anhydrous sodium sulphate was added to a conical flask to cover the bottom by approximately 1 cm. The washed organic phase from the separation vessel was decanted into the conical flask and the conical flask swirled; the previously transparent product solution clouded. The flask was allowed to stand until the organic phase had regained its transparent quality. If a cloudiness remained, more anhydrous sodium sulphate was added with further swirling until the transparent quality returned. The dried product solution was then passed through filter paper and the ethyl acetate evaporated using a RE with warming at 40°C to yield a bright canary yellow solid.

3.3.4 Protocol for cleavage of the side chain protecting OtBu group from Fmoc-L-glu(OtBu)OH-*p*NA

As previously described, the coupling of Fmoc-L-glu(OtBu)OH-*p*NA to the support resin, for subsequent peptide chain extension, was achieved *via* the glutamate side chain; in contrast to a *C*-terminal coupling in conventional SPPS. To prevent linkage of the *p*NA chromophore to the side chain carboxylic acid during the previous stage of synthesis, side-chain OtBu protected glutamate was used. The OtBu protecting group has to be cleaved to allow coupling of the *p*NA derivative glutamate to the support resin. The tertiary OtBu group protecting the glutamate side chain is acid labile and therefore cleavable by addition of TFA with 5% DCM. The presence of a small amount of DCM at this stage prevented the formation of isoprene; a reactive cationic species (156).

A sufficient volume of a 95%TFA:5%DCM solution was added to the solid Fmoc-L-glu(OtBu)OH-*p*NA to achieve dissolution. The reaction solution was then stirred for 30 minutes to allow for complete cleavage of the OtBu group after which TFA was then evaporated using a RE with warming at 40°C. To dry the product (i.e. remove water created from the reaction of the side chain protecting butyl and TFA) washing the residue with 30 ml DCM followed by 30 ml propan-2-ol was carried out. Washing constituted addition of the solvent to the product, gentle swirling and subsequent evaporation of the solvent using a RE with warming at 40°C.

The product Fmoc-L-glu-O-*p*NA was now ready for coupling to the Rink amide resin.

3.3.5 Extension of the peptide chain - overview

As the solid support for extension of the peptide chain, a polystyrene based Rink amide MBHA resin was used (154). The first step in the extension of the peptide chain was the coupling of the side chain deprotected Fmoc-L-glu-O-*p*NA to the resin. The *resin* must first be deprotected by removal of the Fmoc protecting group. The Fmoc group is base labile therefore cleavage was achieved under the basic conditions provided by a 50% morpholine (base): 50% *N*-methyl pyrrolidone (solvent) solution. A 50% morpholine solution was used as morpholine displays very strong basic properties; a 100% solution may lead to deprotonation of incorrect carbons of the Fmoc protecting group resulting in incomplete cleavage from the resin (156). Further, *N*-methyl pyrrolidone causes a desirable expansion of the resin reducing steric hindrance of bulky Fmoc groups thereby allowing sufficient entry to morpholine (156). Other polar solvents such as methanol or ethanol can be used in place of *N*-methyl pyrrolidone though, in practice, they cause an undesirable shrinkage of the resin resulting in less efficient cleavage of the Fmoc protecting groups. The standard manufacturer-recommended procedure of swelling and drying down the resin before use was omitted as in practical terms it has limited value as far as increase in final product yield. All glassware and reagents were thoroughly dried before use so as not to decrease loading efficiency.

Note that in the following text when ‘washing’ is stipulated, all liquid was evacuated under nitrogen, then a volume of the relevant solvent sufficient to cover the reagents by 1-2 cm was added and the reaction vessel placed on a shaker at ambient temperature for 1 minute, following which the solvent was evacuated under nitrogen.

Further, during all reaction stages the minimum volume of solvent was used so to achieve a practicable working volume *but* to maximise reagent concentration. Lastly, all reactions were completed in a siliconised 50 ml reaction vessel.

3.3.6 Protocol for resin deprotection

A sufficient volume of a 50% morpholine: 50% *N*-methyl pyrrolidone was added to solid resin (mass dependent upon scale) at the base of a 50 ml reaction vessel to cover the resin by 1-2 cm and placed on a shaker for 5 minutes. All liquid was evacuated under nitrogen and a further volume of 50% morpholine:50% *N*-methyl pyrrolidone was added to the reaction vessel and shaken for a further 25 minutes. Morpholine was then removed by washing twice with 100% *N*-methyl pyrrolidone since deprotection is more efficient if new base is added following an initial 5 minutes rather than using a single volume of base for the entire 30 minute reaction time (156).

3.3.7 Coupling of Fmoc-L-glu-O-*p*NA to the support resin

The first coupling involves the attachment of the Fmoc-L-glu-*p*NA *via* its side chain to the deprotected resin. For this the coupling agents TBTU and HOBt were used. Simultaneously to coupling, the *N*-terminal Fmoc protecting group of Fmoc-L-glu-*p*NA was cleaved to permit subsequent peptide chain extension. The removal of the Fmoc protecting group from the peptide was completed using a weaker base than used for the resin deprotection; in this instance the more sterically hindered base *di*-isopropyl ethylamine (DIPEA). The relatively strong basic conditions provided by morpholine for resin deprotection were employed to maximise deprotection and correspondingly resin loading capacity. Such conditions are permissible as resin deprotection is completed prior to peptide chain extension so no risk of peptide chain fragmentation exists.

Note that the coupling of the first residue to the resin is referred to as *loading* the resin, it is distinguished from *coupling* of subsequent residues during chain extension simply as it is the first residue in the peptide sequence and is linked to the resin, not another residue. The linkage between the first residue and the resin (when using a Rink amide resin) is through an amide bond as with the linkage between residues of the peptide chain and therefore, in both instances, reaction conditions are identical.

3.3.8 Protocol for loading of the resin with Fmoc-L-glu-O-*p*NA

The solid reagents; Rink amide MBHA resin, Fmoc-L-glu-*p*NA, TBTU, HOBt and *di*isospropyl ethylamine (DIPEA) were added to a 50 ml reaction vessel according to the following molar ratio: 1:3:3:3:9 (note that it is the resin loading capacity used as the molar equivalent here). The resin, Fmoc-L-glu-*p*NA, TBTU, and HOBt solids were added to the reaction vessel followed by a sufficient volume of 100% *N*-methyl pyrrolidone to cover the reagents by 1-2 cm, followed by the relevant volume of DIPEA. The reaction vessel was then placed on a shaker for 1 hour to allow for coupling. Unused reagents were removed by washing through twice with 100% *N*-methyl pyrrolidone. All liquid was evacuated under nitrogen.

Efficiency of loading was tested by ninhydrin assay with the development of a straw yellow colour sought. If additional loading time was required, initially an additional 30 minutes using existing reagents was completed; if further coupling was required all reagents were evacuated under nitrogen, the resin was washed twice with *N*-methyl pyrrolidone and fresh reagents equivalent to half the amounts used initially were added and the reaction vessel then placed on a shaker for 30 minutes. In the rare occurrence that loading had still not been completed it was necessary to *cap* with acetic anhydride (156).

3.3.9 Capping with acetic anhydride

To prevent the synthesis of peptides with missing residues, un-reacted amines (i.e. amino acids that have not coupled the subsequent residue of the peptide sequence and therefore still possess a reactive *N*-terminus) are capped by addition of acetic anhydride to yield an acetylated peptide *N*-terminus. Truncated peptides are easy to remove from the desired product during purification. To these ends a sufficient volume of *di*chloromethane (DCM) was added to cover the reagents by 1-2 cm. Followed by addition of 1 ml of pyridine and 1 ml of acetic anhydride with the reaction vessel placed on a shaker for 10 minutes. The reagents were then washed twice in a 100% solution of *N*-methyl pyrrolidone and the success of capping determined by ninhydrin assay . If capping is successful a straw yellow colour should be seen. If capping is not successful fresh DCM, pyridine and acetic anhydride can be added and a further 10 minutes allowed for capping. Again success of capping is assessed by ninhydrin assay. In the rare occurrence that capping has not proved completely successful no further time for capping should be allowed. Although the conditions required are not harmful to the resin or peptide, further capping is futile and any truncated peptides that may subsequently arise should be separated during product purification. *N*-Terminal capping in this manner was completed at any stage during peptide synthesis where ninhydrin assay proved free amine groups were present (156).

3.3.10 Extension of the peptide chain

Extension of the peptide chain is achieved by successive rounds of deprotection of the *N*-terminal residue of the peptide chain attached followed by coupling of the next Fmoc protected residue. For this, identical protocols, as described for the deprotection and loading (now *coupling*) of the resin, were followed. In the case of synthesis of the longest peptide Ac-DEFQLQ-*p*NA the residues were coupled in the following *C*- to *N*- terminal order: Fmoc-L-Leu, Fmoc-L-Gln, Fmoc-L-Phe, Fmoc-L-Glu, Fmoc-L-Asp, the other four peptides followed the relevant truncation of this order. Note that for cost purposes, with the exception of the *C*-terminal derivative glutamate, all amino acids used in the extension of the peptide chain were not side chain protected; final yields confirmed a low level of impurities despite taking this money saving measure.

3.3.11 Cleavage of Ac-EFQLQ-*p*NA from the resin

Cleavage of the fully synthesised peptide from the resin was achieved under strong acid conditions provided by a 95% TFA: 5% DCM solution along with the cation scavenger *triiso*propylane.

Prior to cleavage from the resin the peptide was thoroughly dried. The reaction vessel was washed through twice with DCM, then once in *di*ethyl ether. Liquid was evacuated under a very gentle flow of nitrogen so as not to create a vortex within the reaction vessel and disturb the dry peptide.

To the dry peptide 200 µl of *triiso*propylane was added followed by a sufficient volume of a 95% TFA: 5% DCM solution to cover the peptide by 1-2 cm. The reaction vessel was then placed on a shaker for 1 hour to allow for complete cleavage. The cleaved peptide, now in solution, was evacuated under a gentle flow of nitrogen into a 250 ml RBF. The cleaved peptide was transparent with a slight yellow tinge. The resin was retained at the base of the reaction vessel and had a burnt orange appearance due to bound salts. All TFA was removed from the cleaved peptide; this was achieved by evaporation using a RE with warming at 40°C for approximately 5 minutes dependent upon starting volume. An oily yellow layer formed around the middle of the RBF; this is the peptide along with small amounts of TFA and DCM.

3.3.12 Protocol for peptide precipitation

It is necessary to precipitate the peptide from the oily layer by the use of ether. To the oily liquid, 20 ml of *di*ethyl ether was added and the RBF swirled. A thin layer of precipitate will immediately be observed. To protect the peptide product from UV degradation, the RBF was covered in aluminium foil and allowed to stand in a fume cupboard for 1 hour to allow for further precipitation of product.

A thin layer of white precipitate with a slight yellow tinge was observed around the middle of the RBF. The *di*ethyl ether at the base of the reaction vessel contains TFA, DCM and some impurities. The majority of ether was then carefully decanted off with the remainder being removed under vacuum by attachment of the RBF directly to a vacuum pump for approximately 5 minutes. Before attachment to the vacuum pump a light tapping of the sides of the RBF to cause any solid to fall to the base will prevent spurting of any peptide if any oily residue remains between the solid and the side of the RBF. The crude peptide was then stored, away from light at 4°C until required for purification (156).

3.4 Analysis and purification of the synthetic peptides

3.4.1 Sample analysis

Purity of the synthetic peptides was determined by analytical reverse phase chromatography using a Phenomenex C12 Jupiter column (250 x 4.6 mm) 90 Å bead column connected to a Waters 6006 HPLC system. Sample separation was achieved by gradient elution from 0% acetonitrile (+0.1% TFA) in water to 100% acetonitrile (+0.1% TFA) over 100 ml at a flow rate of 2 ml/min to allow for good species separation. Due to the highly hydrophobic nature of both peptides, crude peptide was prepared by dissolution of 0.5 mg of crude product in 200 µl DMSO followed by addition of analytical grade water sufficient to just induce precipitation of the product and the formation of a fine suspension. This suspension was then loaded directly onto the column. Eluting peptides were monitored by UV absorption at 216 nm and relevant peaks were analysed by mass spectrometry in positive phase using a VG Quattro II mass spectrometer. Samples were prepared for mass spectrometry by dilution 10 fold in a 50% acetonitrile: water solution.

3.4.2 Sample purification

Purification of all synthetic peptides was achieved in a similar manner as for analysis but with the following exceptions; a Phenomenex C12 Jupiter column (250 x 10 mm) 90 Å bead size reverse phase column was used; and a gradient of 20% to 65% of acetonitrile (+0.1% TFA) in water over 100 ml at a flow rate of 5 ml/min was used. The crude sample was prepared by dissolution of 10 mg crude peptide in 100 µl DMSO with water added to form a suspension as with sample preparation for analytical chromatography; typically a final volume of 200 µl would be achieved. The crude peptide was then purified in a batch-wise fashion; 10 mg of crude peptide per run (156).

3.4.3 Peptide storage

With the peptide fresh from the column in approximately 40% (dependent upon peptide length) acetonitrile (+0.1% TFA), the sample was transferred to a 250 ml RBF, placed in liquid nitrogen and spun by hand until the sample froze. The frozen sample was then freeze dried overnight or until the sample was visibly solid but fluffy in appearance. The dry, fully purified sample was then stored in the dark at 4°C until use (156).

3.4.4 Assay of 3C protease with chromogenic peptide

To assay qualitatively for 3C protease activity during and after protein purification, 0.1 mg of Ac-EFQLQ-*p*NA was dissolved in 100 µl of DMSO; a volume equivalent to 3 nmoles of SV3CP was added followed by the assay buffer: 100 mM TRIS pH 8.5 5 mM *β*-mercaptoethanol, to a final volume of 500 µl. The sample was then incubated at 37°C with absorbance at 405 nm (i.e. development of a yellow colour) measured spectrophotometrically over a 10 minute period. If the SV3CP is satisfactorily active a strong yellow colour will develop after only 1-2 minutes.

See results section for a brief description of the assay conditions used for the preliminary peptide specificity and inhibition assays.

3.5 Basis of design of a highly potent inhibitor of SV3CP

Based upon the preliminary chromogenic peptide work the peptide resulting in the maximum rate of cleavage possessed the sequence: Ac-EFQLQ-*p*NA. It was therefore reasoned that an inhibitor based upon this sequence would have a high binding specificity for the SV3CP active site cleft. Peptidyl inhibitors in which the scissile peptide bond is substituted by a Michael acceptor were initially designed and tested on the cysteine protease *papain* (157). Adopting this approach an inhibitor was designed where an *ethyl-ester* extension was coupled to the *C*-terminal glutamate of the peptide sequence EFQLQ (*Figure 3.7*). The ethyl-ester moiety behaves as a Michael type acceptor undergoing nucleophilic attack by the active site cysteine-thiol resulting in the inhibitor becoming covalently and irreversibly bound to the active site cysteine (158). Such inhibitors are sometimes referred to as 'suicide' inhibitors. An ethyl-ester group is a relatively weak electrophile, therefore it must be in close proximity to the active site cysteine. It was reasoned that the peptide portion of the inhibitor would direct the Michael acceptor group to the active site cleft causing the inhibitor to bind at the correct position at the active site and to be specific for the SV3CP active site cysteine only. The inhibitor was synthesised using a combination of standard solid phase peptide synthesis and synthetic chemistry techniques.

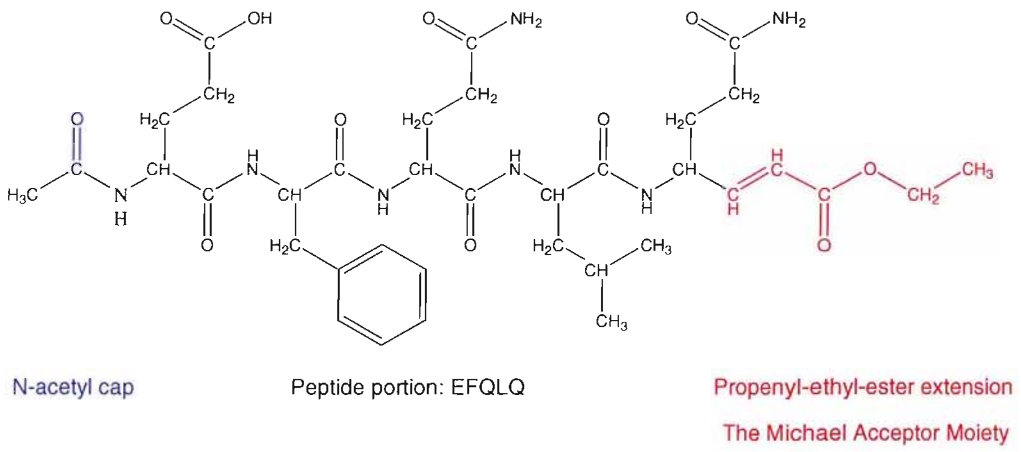


Figure 3.7: Structure of the Michael acceptor peptidyl inhibitor (MAPI) designed and synthesised to be both highly potent and highly specific for the SV3CP.
The inhibitor covalently and irreversibly binds to the active site cysteine of the SV3CP to produce the covalent adduct, rendering the protease entirely inactive (*Figure 3.8*).

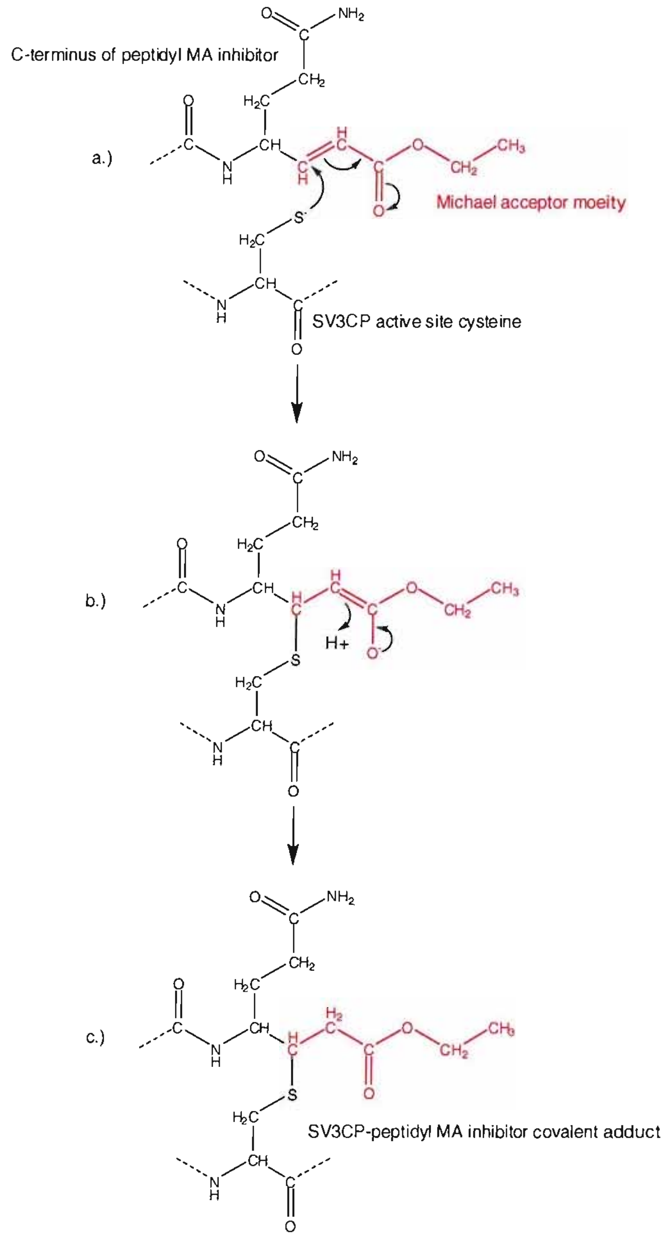

C-terminus of peptidyl MA inhibitor
a.)
Michael acceptor moeity
SV3CP active site cysteine
b.)
H+
c.)
SV3CP-peptidyl MA inhibitor covalent adduct

***Figure 3.8: Mode of action of MAPI**. a.) Nucleophilic attack of the C=C of the Michael acceptor moiety by the SV3CP active site cysteine Cys-139. b.) Intermediate species. c.) the SV3CP-peptidyl MA covalent adduct.*

3.6 Synthesis overview of MAPI

The following scheme provides an overview of the approach taken in synthesis of MAPI (40, 156).

Generation of the Weinreb amide

Starting with *N*-terminal Boc, and side chain *trityl* -protected glutamine; the *C*- terminus hydroxyl is de-protonated by addition of the base TEA to leave O^- : the *acyl intermediate*. The glutamine acyl intermediate is then activated by addition of BOP. This allows N,O,*di*methylhydroxylamine to be coupled to the *C*- terminus to yield the *Weinreb amide* (alternatively termed the *N-methoxy-N-methylamide* group) *(Figure 3.9: b.)*. See Section 3.7 for method of synthesis.

↓

Reduction to the aldehyde

The Wienreb amide is reduced to form an aldehyde *(Figure 3.9 c.)* by reaction with DIBAL; a strong reducing agent. It is essential to work quickly at this stage to prevent cyclisation of the aldehyde and species racemisation. See section 3.8 for method if synthesis. See Section 3.8 for method of synthesis.

↓

The Wittig-Horner reaction

The Wittig-Horner reaction *(Figure 3.9 d.)* involves converting the aldehyde into the final product*, the *glutamine proponyl-ethyl-ester,* by use of a *phosphonoacetate* that reacts with the aldehyde to add the *proponyl-ethyl-ester* group. This is completed in the presence of a base: *sodium bis(trimethylsilyl)amide* an *alkali metal salt* that activates the phosphonoacetate for coupling. See Section 3.9 for method of synthesis.

*The "final product" being modified glutamine, not the completed MAPI.

↓

De-protection of N-terminus of glutamine-propenyl-ethyl-ester by cleavage of the Boc protecting group

Attached to the *N*-terminus of the modified glutamine is a Boc protecting group, this is cleaved in a 20% solution of TFA in DCM to provide the free NH_2 group (*Figure 3.9: e.*)

for coupling to the remainder of the peptide: EFQLQ. See Section 3.10 for method of synthesis.

↓

Coupling of de-protected glutamine-propenyl-ethyl-ester to the peptide Ac-EFQL

Coupling (*Figure 3.9: f.*) is achieved in solution, under basic conditions (pH 8; by addition of sufficient DIPEA) in a solution of 50:50; DCM:DMF with DIC as the coupling agent. See section 3.10 for method of synthesis. See Section 3.10 for method of synthesis.

↓

Final stage: cleavage of side chain protecting groups

Glutamine possesses a side chain protecting trityl group, and, glutamate a tertiary butyl group, these are both cleaved in 70% TFA in DCM (*Figure 3.9 g.*). See Section 3.10 for method of synthesis.

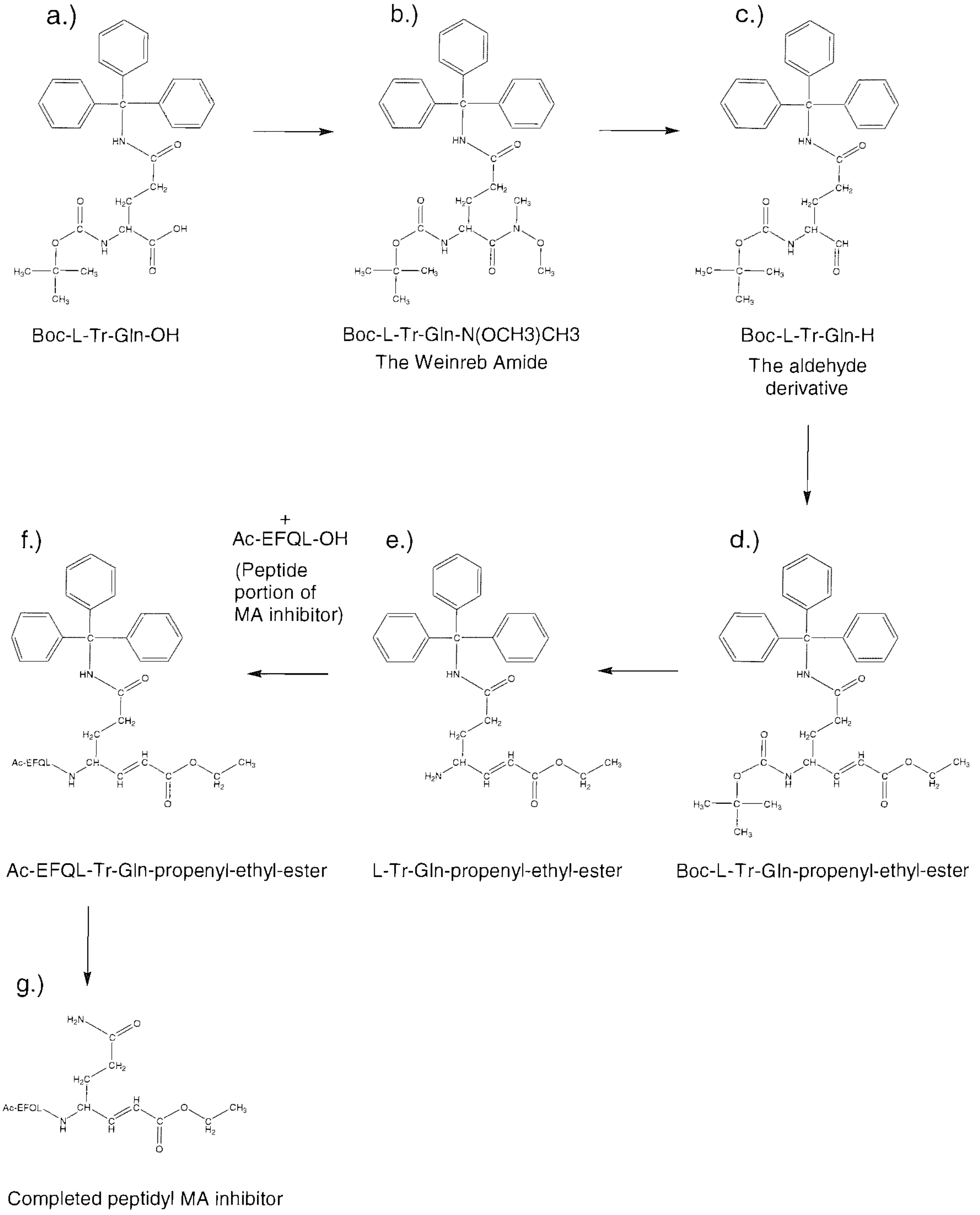

a.)
Boc-L-Tr-Gln-OH
b.)
Boc-L-Tr-Gln-N(OCH3)CH3
The Weinreb Amide
c.)
Boc-L-Tr-Gln-H
The aldehyde derivative
d.)
Boc-L-Tr-Gln-propenyl-ethyl-ester
e.)
L-Tr-Gln-propenyl-ethyl-ester
+
Ac-EFQL-OH
(Peptide portion of MA inhibitor)
f.)
Ac-EFQL-Tr-Gln-propenyl-ethyl-ester
g.)
Completed peptidyl MA inhibitor

Figure 3.9: Schematic representation of the synthesis of the peptidyl MA inhibitor of SV3CP.

3.7 Generation of the Weinreb amide glutamine derivative

3.7.1 Background chemistry

This is an essential step to make the *C*-terminal carbon of the glutamine reactive to accept a nucleophile so that the aldehyde may be generated. In its normal state i.e. as a carboxylic acid, the *C*-terminal carboxyl of glutamine is quite resistance to nucleophilic attack as the oscillation of negative charge between the two carboxylic acid oxygen atoms deactivates the positive charge of the carbon nucleus. To increase the δ^+ of the carbon, one of the oxygen groups has to be replaced with an electron withdrawing group, such as an ester or, as used here, a hydroxylamine. Both serve the same purpose. The *C*-terminal carbon then becomes quite δ^+ and can be converted in to the necessary aldehyde during the following step (160).

3.7.2 Experimental procedure

3.7.2.1 Synthesis

The following synthesis was completed in a non-stoppered 250 ml RBF. Boc-L-Tr-Gln-OH (5 mmoles) were dissolved in 50 ml of DCM followed by addition of the base TEA (1 eq)[a], then immediately addition BOP (1 eq)[a]. The reaction mixture was then stirred at RTP for 15 minutes. Following 15 minutes reaction time *N,O,dimethylhydroxylamine hydrochloride* (1 eq)[a] was added (the Weinreb amide), and further TEA (1 eq)[a], followed by stirring for 2 hours at RTP (*Figure 3.10*). Additional TEA is required at this stage to neutralise HCl produced by *N,O,dimethylhydroxylamine hydrochloride*. Good practice at this stage is to run a mass spectrum of a small amount of the *N,O,dimethylhydroxylamine hydrocholride* before use, as this reagent is often supplied 10-30% impure, where the impurity is *N,O,dimethylhydroxylamine hydrochloride* minus the hydroxyl group, high purity is essential to maximise final product yield. Following the 2 hour reaction time, remaining starting reagents were checked for by TLC and mass spectrometry. As no starting reagents remained the need for an intermediate purification step was removed and therefore the work-up was begun directly (40, 156).

[a] Relative to the starting number of moles of Boc-L-Tr-Gln-OH

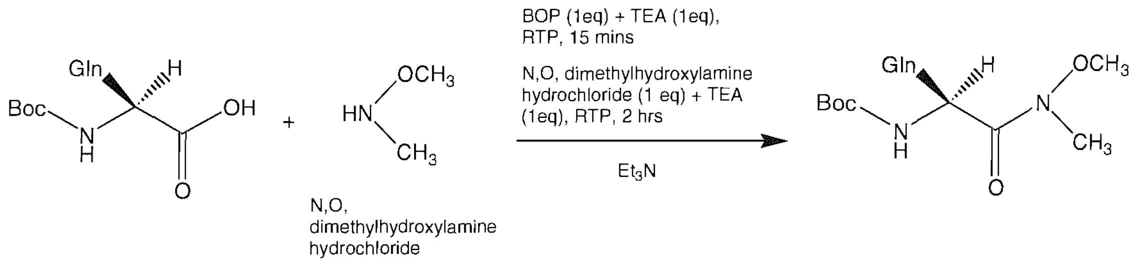


Figure 3.10: Reaction conditions required for the generation of the Weinreb amide derivative glutamine.

3.7.2.2 The Work-up

DCM was evaporated from the reaction solution using a RE with warming at 40°C. The solid product was then dissolved in 200 ml of ethyl acetate and the solution transferred to a 500 ml separating vessel. It is for convenience that the product was re-dissolved in ethyl acetate; during the following wash steps ethyl acetate forms an immiscible layer on top of the aqueous phase in the separating vessel; in contrast DCM has a higher density than water and would lie beneath the aqueous phase making for a more difficult workup. The solution was then washed once in 200 ml of a 1 M HCl solution. If an emulsion formed between the organic (upper) and aqueous (lower) phases the solutions were allowed to settle to allow this to disperse. HCl neutralises the base TEA drawing the resultant HCl salt of TEA into the aqueous phase and out of the organic phase containing the product. The use of ethyl acetate as the organic solvent has a second benefit during this wash stage as DCM would solvate HCl to a higher degree. By using ethyl acetate, HCl will stay mainly in the aqueous phase reducing the likelihood of cleaving the trityl group protecting the side chain of the Weinreb amide glutamine derivative. A second wash of the organic phase was completed but with 200 ml of a super-saturated $NaHCO_3$ solution in analytical grade water. $NaHCO_3$ neutralises residual HCl from the previous wash and generates the sodium salt of the carboxyl terminus of any unreacted *Boc-L-Tr-OH* causing its precipitation out of the organic phase and into the aqueous phase. The product, dissolved in ethyl acetate, was then dried over solid

Na_2SO_4 followed by evaporation of the ethyl acetate on a RE with warming at 40°C. The solid product was weighed to determine yield; a yield of 80% or above should be achievable. Mass spectrometry was then employed to check for the correct product mass of 531.64 g. Purity can also be checked by TLC, though this is a somewhat unnecessary step as the product should be quite pure and there should be no need for any purification steps before continuing to generation of the aldehyde (156).

3.8 Reduction of the Weinreb amide glutamine to the aldehyde derivative

3.8.1 Background Chemistry

DIBAL (*diiso*butyl aluminium hydride) belongs to a class of reducing agents named the metal hydrides capable of reducing, amongst other chemical moieties, a Weinreb amide to an aldehyde. DIBAL forms a stable intermediate *via* interaction of its aluminium to the two oxygen groups of the Weinreb amide extension. Upon decomposition, induced under acidic conditions provided by addition of HCl, the intermediate species collapses to yield the aldehyde derivative glutamine (*Figure 3.11*). It is important that DIBAL is used in preference to more potent metal hydrides to prevent further reduction of the desired aldehyde to a primary alcohol. For example: each mole of the strongly reducing lithium aluminium hydride ($LiAlH_4$) will reduce four moles of Weinreb amide as each of the four H atoms of $LiAlH_4$ is chemically active. In contrast, the more sterically hindered DIBAL will reduce equimolar amounts of the Weinreb amide as it possesses just a single active H. It is clear to see that, in equimolar ratios, DIBAL can only reduce a Weinreb amide to the aldehyde and not to the primary alcohol (156).

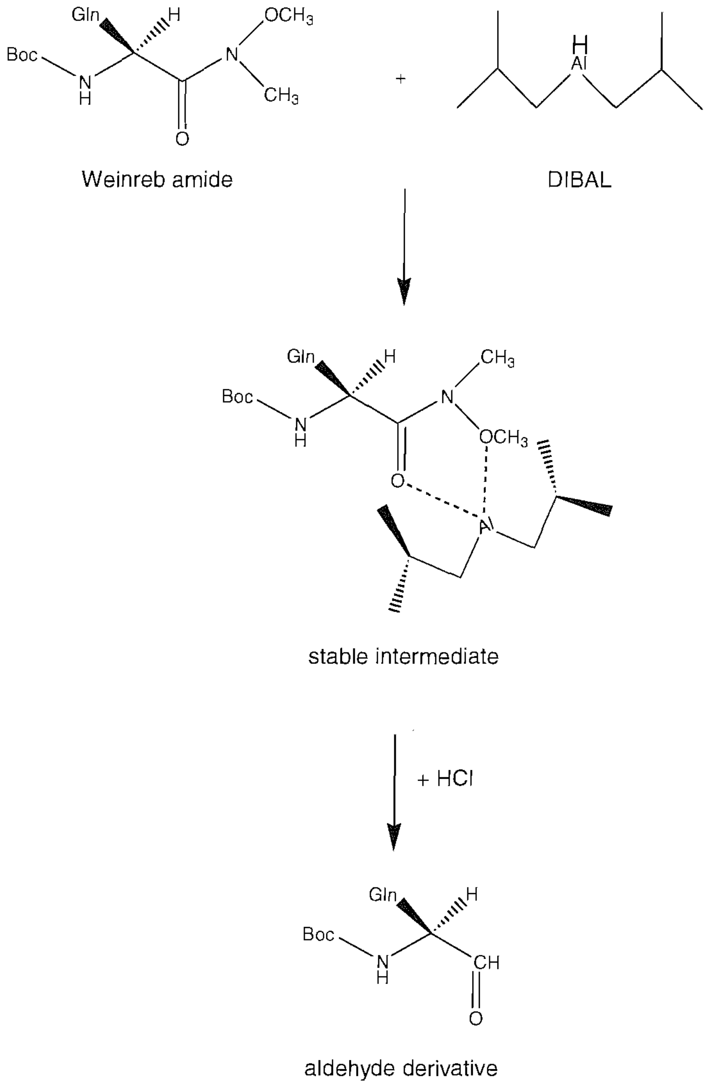


Figure 3.11: Reduction of the Weinreb amide derivative glutamine to the aldehyde by the reducing metal hydride DIBAL. *Addition of HCl acid provides the acidic conditions necessary for the collapse of the intermediate species conjugate to yield the aldehyde derivative.*

3.8.2 Experimental procedure

3.8.2.1 Synthesis

The following synthesis was completed in a 250 ml RBF under slight positive pressure of argon (any inert gas is suitable). Reagents were added *via* canula, and any exceptions are noted in the text.

The Weinreb amide glutamine derivative produced in the previous step was dissolved in 40 ml of THF at RTP in an un-stoppered RBF. Following dissolution the RBF was stoppered with a Superseal style stopper. Through the Superseal stopper were inserted: a canula (the gas canula; the inert gas was always supplied *via* this route) attached to a supply of inert gas (of adjustable pressure); a canula sealed by Parafilm to a balloon[†]; and a wide bore hypodermic needle. Air was evacuated from the RBF for 1 minute under a low pressure of inert gas provided *via* the gas canula. The wide bore hypodermic simply provides an exit for the displaced air and is therefore removed following the 1 minute period of evacuation. Using the minimum necessary flow of inert gas the balloon was inflated until just pert. The balloon remained attached throughout the generation of the aldehyde as it serves two purposes; by providing a slight positive pressure of an inert gas therefore preventing air from re-entering the RBF and allowing room for gases generated during the reaction to fill the balloon, hence it is essential not to fully inflate the balloon initially. The RBF and balloon assembly was lowered into a dry ice-acetone bath to maintain a temperature of -78°C throughout.

The solution in the RBF was allowed to cool to -78°C; this can be judged by observing the acetone of the dry-ice-acetone bath: when it stops boiling the RBF and its contents will be close to -78°C. The DIBAL (2.5 eq)[a] was added dropwise *via* canula to the Weinreb amide and the reaction solution stirred for 4 hours (*Figure 3.12*). During the first minutes of the reaction hydrogen will be given off and bubbling will be seen. When bubbles cease the reaction would appear to be over. However although the absence of bubbles is suggestive of a completed reaction, and therefore generation of the aldehyde glutamine derivative, it is necessary to allow the full 4 hour reaction time to ensure the reaction reaches completeness. Following the full reaction time, the reaction was terminated by addition of 2 ml of methanol. Methanol causes degradation of uncomplexed DIBAL. After 1 minute to allow for complete degradation of DIBAL, HCl was added *via* canula to a final concentration of 0.1 M to release the DIBAL and yield the aldehyde derivative glutamine. The RBF was then removed from the dry-ice-acetone bath and allowed to warm to room temperature (40, 156).

[a] Relative to the starting number of moles of Boc-L-Tr-Gln-N(OCH_3)CH_3.

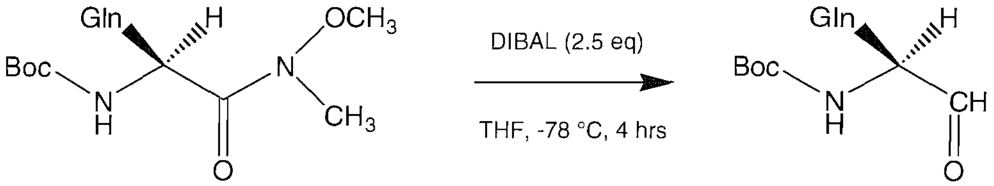


Figure 3.12: Reaction conditions required for the generation of the aldehyde derivative glutamine.

3.8.2.2 The Work-up

The work-up was completed at RTP, accordingly the RBF was un-stoppered and the balloon discarded. From this point onwards it was imperative to work as quickly as possible with the aldehyde to prevent further reduction to the primary alcohol derivative.

On removal from the dry ice-acetone bath the solution will form a suspension. The suspension was brought back into solution by addition of 150 ml of *di*ethyl ether. The solution was transferred to a 500 ml separating vessel (of the non-equilibrating type) and washed three times with 100 ml of a 0.1 M solution of HCl. It should be appreciated that at this concentration of HCl there is no risk of cleaving the trityl group protecting the side chain of the aldehyde derivative glutamine. The solution was then washed with a half-saturated solution of $NaHCO_3$ and a final wash with tap water. After washing, a white layer may be observed between the organic and aqueous phases; this is DIBAL, still in complex with the aldehyde, which was not successfully released after the addition of the HCl. Since this layer contains some product it should considered as the organic phase and treated accordingly. Note that each wash step involves vigorous shaking to form an emulsion, followed by a stationary period to allow the emulsion to settle into the organic (upper) and the aqueous (lower) phases. The lower aqueous phase will contain unreacted species and therefore waste; the upper organic phase will contain the aldehyde glutamine derivative product. Accordingly, the lower aqueous phase was decanted and discarded following each wash step. The final organic phase was dried over a sufficient mass of solid anhydrous $MgSO_4$ and filtered. It is important to note that the organic phase should not be left to dry over night as employed during generation of the Weinreb amide; $MgSO_4$ is slightly acidic and will cause the aldehyde to racimise over a prolonged period. The *di*ethyl ether and residual toluene (DIBAL is usually supplied in toluene) were evaporated on a RE with warming at 40°C. Due to the high boiling point of 100°C of toluene, an oil of the product will form. The oil was placed at 4°C to form a jelly. The jelly can be frozen at -20°C until further use if necessary. To remove toluene completely and to achieve a solid of the product, a 250 ml RBF containing the product oil was connected directly to a diaphragm pump, whilst warming was supplied by placing the RBF in a water bath at 40°C. During this stage care was taken to prevent the oil from heating above 60°C (oil will behave as a heat sink and therefore its temperature may rise above the 40°C of the water bath) leading to degradation of the product. Finally, the product was subject to mass spectrometry to check for correct M_r of 473.4 g. Note that the mass spectrum is likely to contain a peak corresponding to the trityl group alone at 243.2 g; this is due to fragmentation of the product during mass spectrometry *not* due to cleavage during synthesis of the aldehyde. Also, peaks are likely to be seen corresponding to the dimer of the aldehyde carrying a single charge at 945.8 g. There is no need for any further purification steps before moving on to generate the propenyl-ethyl-ester by *via* the Wittig-Horner reaction (156).

3.9 The Wittig-Horner Reaction

3.9.1 Background Chemistry

The Wittig-Horner reaction is an adaptation of the original Wittig reaction where the alkylating phosphonium ylide is replaced by a *phosphonate stabilised carbanion* to generate E-alkenes in favour of the Z-alkenes generated with the Wittig reaction. This stage involves the conversion of the aldehyde derivative glutamine to an unsaturated ester by reaction with a phosphonate stabilised carbanion. The deprotonation of the phosphonate *triethyl phophonoacetate* under the strongly basic conditions provided by *sodium bis(trimethylsilyl)amide* generates the stabilised *carbanion*. The carbanion then reacts with the aldehyde derivative glutamine proceeding *via* an intermediate complex to ultimately yield the E-alkene derivative propenyl ethyl ester glutmamine (*Figure 3.13*) (156).

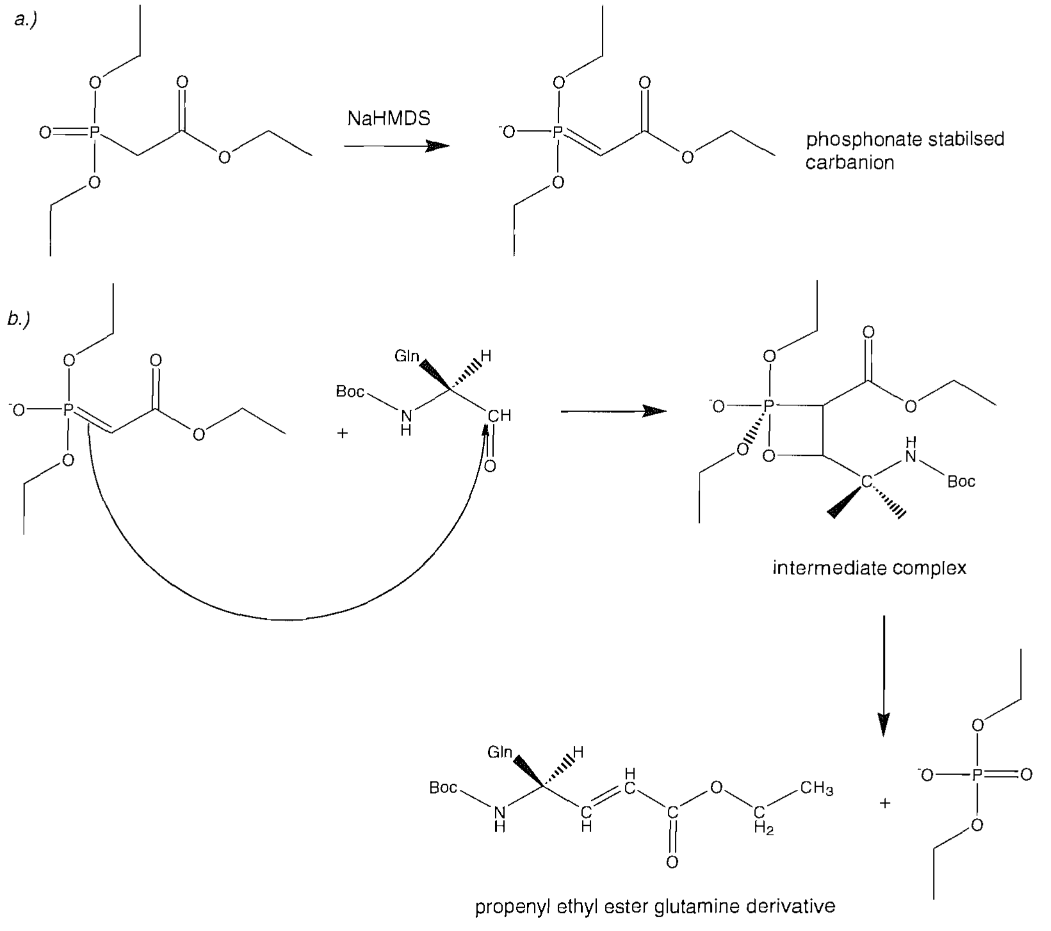

a.)
NaHMDS
phosphonate stabilsed carbanion
b.)
Gln
H
Boc
N
H
CH
O
+
intermediate complex
Boc
propenyl ethyl ester glutamine derivative
CH_3
H_2

Figure 3.13: The Wittig-Horner reaction. *a.) Generation of the phosphonate stabilised carbanion by the deprotonation of the phosphonate triethyl phophonoacetate by the base sodium bis(trimethylsilyl)amide (NaHMDS). b.) Reaction of the phosphonate stabilised carbanion species with the aldehyde derivative glutamine via the four-membered cyclic intermediate to yield the E-alkene propenyl-ethyl-ester glutamine derivative.*

3.9.2 Experimental procedure

3.9.2.1 Synthesis

The following synthesis was completed in a 250 ml RBF under slight positive pressure of an inert gas and at -78°C, this can be achieved by following the apparatus set-up instructions in the experimental section detailing the synthesis of the aldehyde derivative of the glutamine. Reagents were added *via* canula, any exceptions to this are noted in the text.

An empty RBF was allowed to cool to -78°C in a dry ice-acetone bath. The *triethyl phosphonoacetate* (1 eq)[a] was added to the RBF followed by 20 ml of THF and stirred for 5 minutes to cool to -78°C. Following dissolution, *sodium bis(trimethylsilyl)amide* (1 eq)[a], the *triethyl phosphonoacetate* activating agent, was added and the reaction mixture stirred for 20 minutes. This period of stirring will allow for deprotonation and subsequent activation of the *triethyl phosphonoacetate* by generation of the stabilised carbanion. It is important that the reaction solution remains colourless during this stage; a yellow colour is indicative of the reaction proceeding too speedily; keeping the reaction solution at -78°C helps to slow the activation of the *triethyl phosphonoacetate*. The aldehyde glutamine derivative (*Boc-L-Tr-Gln-OH*) synthesised in the previous step was dissolved in 20 ml of chilled THF and then added to the solution in the RBF. The reaction was then allowed to proceed for 2 hours at -78°C with stirring, during which time the balloon filled with inert gas was topped up to maintain a positive pressure (gas added *via* canula) (*Figure 3.14*). Following 2 hours reaction time the RBF was removed from the dry ice-acetone bath and placed in a water-ice bath to allow warming to 0°C. A slight yellow colour may develop during this period due to deprotonation of the silane ion, a by-product of deprotonation of *triethyl phosphonoacetate* by *sodium bis(trimethylsilyl) amide*, but is not of concern (40, 156).

[a] Relative to the starting number of moles of Boc-L-Tr-Gln-H

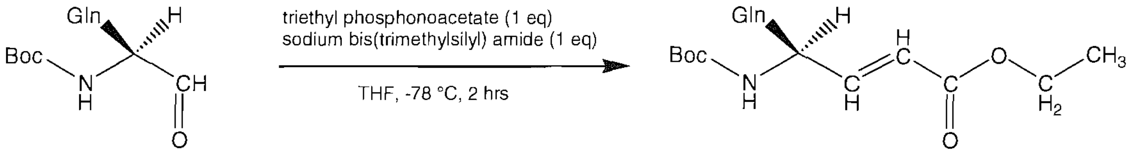

Gln
H
Boc
N
H
CH
O
triethyl phosphonoacetate (1 eq)
sodium bis(trimethylsilyl) amide (1 eq)
THF, -78 °C, 2 hrs
Gln
H
Boc
N
H
H
C
C
H
C
O
O
C
H2
CH3

Figure 3.14: Reaction conditions required for the generation of the propenyl ethyl ester derivative glutamine.

3.9.2.2 The work-up

The work-up was completed at RTP. The reaction solution was decanted into a 500 ml separating vessel, followed by addition of 100 ml of a 0.5 M HCl solution and 100 ml of a mixture of equal volumes of *ethyl acetate* and *hexanes*. The separating vessel was then shaken vigorously to form an emulsion followed by a stationary period to allow the separation into the aqueous (lower) and organic (upper) phases. The organic phase should lose any yellow colour gained in the final step of synthesis due to reprotonation of the *silane ion* by washing in HCl. The organic phase contains the washed product and therefore was retained The lower aqueous phase contains impurities (unreacted species), but also may contain some product, so a second wash of just the aqueous phase was undertaken with 100 ml of a solution of equal volumes of ethyl acetate and hexanes. THF is very slightly miscible with water and since it is used as the solvent during the reaction stage it will form part of the organic phase during the wash cycles. This can lead to a small amount of product being transferred into the aqueous phase during washing. The organic phases from the first and second washes were combined, dried over solid Na_2SO_4 and then filtered. The solvent was evaporated from the product by use of a RE with warming at 40°C. Evaporation of the solvent may take some time as the boiling point of ethyl acetate and hexanes in combination is approximately 65°C. A pale yellow oil formed on evaporation of the solvent, at which point heating was stopped to prevent degradation of the product. The oil containing the product is quite stable and can be stored at -20°C until further use. Purity of the product *Boc-L-Tr-Gln-propenyl ethyl ester* was determined using TLC. Mass spectrometry should confirm the correct M_r of 542.7 g with no other significant peaks, although small peaks at the following masses[†] are likely to be seen; -100 corresponding to product minus Boc protecting group; -56 fragmented Boc group; -243 minus trityl protecting group; and -29 resulting from fragmentation at the ester bond. These additional peaks are due to fragmentation during mass spectrometry and are not as a result of species generated during synthesis of the propenyl-ethyl-ester derivative glutamine. Due to the reluctance of Boc-L-Tr-Gln-propenyl-ethyl-ester to 'fly' during mass spectrometry a high cone voltage is required, this exacerbates fragementation of the parent ion: Boc-L-Tr-Gln-propenyl-ethyl-ester $^+$, but is not of concern. The yield of Boc-L-Tr-Gln-propenyl-ethyl-ester from Boc-L-Tr-Gln-H should be at least 80%. Purification was not necessary before progression to coupling the propenyl-ethyl-ester glutamine derivative with the peptide Ac-EFQL to form the final inhibitor *Ac-EFQLQ-propenyl-ethyl-ester* (40, 156).

†subtracted from the mass of 542.7 corresponding to the product: Boc-L-Tr-Gln-propenyl-ethyl-ester.

3.10 To couple Boc-L-Tr-Gln-propenyl ethyl ester to the peptide Ac-EFQL

The coupling of the Boc-L-Tr-Gln-propenyl-ethyl-ester to the peptide Ac-EFQL was achieved in three stages: cleavage of the Boc group protecting the *N*-terminus of Boc-L-Tr-Gln-propenyl-ethyl ester; solution phase coupling of the deprotected L-Tr-Gln-propenyl-ethyl-ester to the peptide Ac-EFQL; and finally the cleavage of the *C*-terminal trityl side chain protecting group of Ac-EFQLQ-propenyl-ethyl-ester.

3.10.1 To cleave the *N*-terminal Boc protecting group of Boc-L-Tr-Gln-propenyl-ethyl ester

The oil containing Boc-L-Tr-Gln-propenyl-ethyl-ester was thawed and allowed to reach RT before dissolution in 50 ml of DCM. A sufficient volume of TFA to make 20% v/v of the final solution was added then the solution stirred at RTP for 10 minutes (*Figure 3.15*). A strong yellow colour will be seen, this is due to a small proportion of the trityl side chain protecting groups being cleaved in the acidic conditions provided by addition of TFA yielding the undesired trityl cation that is quite yellow in colour. However, at

20% TFA the majority of side chains remain protected. Following stirring, solvent was evaporated using a RE with warming at 40°C. A residual amount of TFA may remain, and is quite easily identifiable by smell. To remove residual TFA, 50 ml DCM was added, the RBF swirled then the solvent evaporated as before. The DCM will gradually draw off the remaining TFA. If TFA is still present the dissolution of product in DCM can be repeated. To precipitate the product, 20 ml *di*ethyl ether was added and the RBF swirled; the propenyl-ethyl-ester derivative glutamine minus its Boc protecting group *(L-Tr-Gln-propenyl-ethyl-ester)* will form a yellow precipitant. The *di*ethyl ether was then evaporated by use of a RE with warming at 40°C; both TFA and any cleaved trityl groups will solvate in ether therefore both species should be drawn off along with the diethyl ether. The yellow L-Tr-Gln propenyl-ethyl-ester was then re-dissolved in a further 20 ml of *di*ethyl ether, the RBF swirled and *di*ethyl ether evaporated as before. The resultant solid should now be off-white in colour as more of the residual TFA and trityl cation are drawn off with the evaporating diethyl ether. The L-Tr-Gln-propenyl-ethyl-ester was then dissolved in 20 ml of acetonitrile, the RBF swirled and solvent evaporated on a RE with warming at 40°C to leave a white solid; the pure L-Tr-Gln-propenyl-ethyl-ester. Success of cleavage of the *N*-terminal protecting Boc group can be judged by ninhydrin assay. Mass spectrometry can be used at this stage to verify the cleavage of the Boc protecting group but it should be noted that L-Tr-Gln-propenyl-ethyl-ester will exist at this stage as the TFA salt and will therefore have a M_r of 561.7 g not 442.7 g ; the correct predicted M_r.

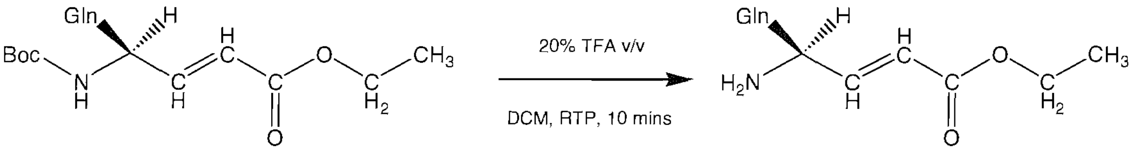


Figure 3.15: Reaction scheme for cleavage of the N-terminal Boc protecting group of Boc-L-Tr-Gln-propenyl-ethyl-ester.

3.10.2 Coupling procedure

In a 100 ml RBF, 400 mg of L-Tr-Gln-propenyl-ethyl-ester was dissolved in 20 ml of DCM. A sufficient volume of DIPEA (~9 eq)[a] was added to achieve a pH of 8; the basic conditions required for Fmoc peptide chemistry. Note the pH of an organic solvent solution can be determined by dropping a small volume onto damp pH paper. On addition of the DIPEA fumes may be observed, due to the neutralisation of the L-Tr-Gln-propenyl-ethyl-ester TFA salt. If the solution has a cloudy appearance it can be clarified by adding a small volume of DMF. Separately, Ac-EFQL (1 eq)[a] was dissolved in 2 ml of a solution of DCM and DMF in equal volumes then added the to the *L-Tr-Gln-propenyl-ethyl ester* followed by DIC (2 eq)[a]. The reaction solution was then stirred at RT for 4 hours to allow coupling (*Figure 3.16*). Following coupling, solvent was evaporated by use of a RE with warming at 40°C until reduced to approximately 5 ml of a viscous straw yellow solution; a solid is prevented from forming due to residual DMF. To this viscous solution 50 ml of *di*ethyl ether were added causing immediate precipitation of the product: *Ac-EFQLQ-propenyl-ethyl-ester* albeit with the *C*-terminal glutamine side chain still protected by a trityl protecting group. The solution should be left for 30 minutes to maximise yield. To recover the product, the contents of the RBF were transferred to an empty PD10 column; the *di*ethyl ether and DMF will pass through the glass sinter of the PD10 column to leave a gel of the product. The residual ether was allowed to evaporate from the gel by leaving at RT for 1 hour.

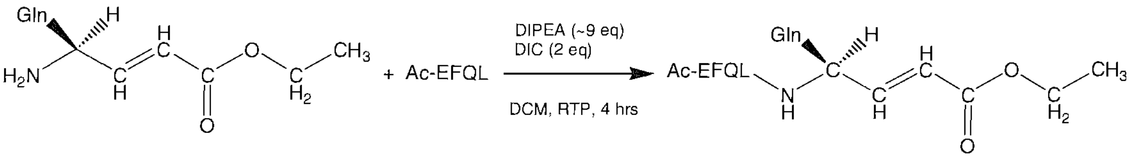


Figure 3.16: Reaction conditions for coupling of L-Tr-Gln-propenyl ethyl ester to the peptide Ac-EFQL.

3.10.3 To cleave side chain protecting groups

The gel of the product *Ac-EFQLQ--propenyl-ethyl-ester* was transferred to a 100 ml RBF. The side chain protecting groups were then cleaved in 20 ml of a solution of 70% TFA in DCM at RTP with stirring for 1 hour. To scavenge the cleaved trityl cation, 300 µl of *tri-isopropylsilane* was added during this cleavage reaction. The solvent was then evaporated using a RE with warming at 40°C to leave an oil of the product. The product was triturated from the oil to yield a gel by addition of 20 ml of *di*ethyl ether. The gel was then dissolved in 20 ml of a solution of equal volumes of acetonitrile and analytical grade water and dried to form a white solid of the final product: *Ac-EFQLQ-propenyl-ethyl-ester* by freeze drying. Mass spectrometry should confirm a M_r of 759, although the sample is highly likely to be impure with peaks at 902 and 577 corresponding to the product with side chain protecting groups remaining un-cleaved, and the uncoupled peptide Ac-EFQL respectively. Purification of the final product from impurities was therefore necessary and was achieved by reverse phase chromatography (see *Section 4.3.2*).

3.11 Analysis and purification of Ac-EFQLQ-propenyl-ethyl-ester

No more than 10 mg of the crude product was dissolved in 100 µl of DMSO then immediately before loading onto a Phenomenex C12 Jupiter column (250 x 10 mm) 90 Å bead size reverse phase column, 100 µl of analytical grade water was added or a sufficient volume to just cause a suspension to form. A gradient was run from 20 to 55% of acetonitrile (+0.1% TFA) in water over 25 minutes at a flow-rate of 4 ml/min. Purity of the final product was determined by mass spectrometry with a single major peak sought in the mass spectra corresponding to a M_r of 759.

4.0 Results

4.1 Expression and purification of SV3CP

4.1.1 Expression of native SV3CP from plasmid construct: pSV3C

A pET based expression vector (161) had previously been constructed by Dr. M. Sarwar, University of Southampton, by insertion of the gene encoding SV3CP (kindly provided by Dr. P. Lambden and Prof. I. Clarke of the Molecular Microbiology and Infection Group, University of Southampton), along with flanking regions 5' and 3', between the restriction sites *Nde*1 and *Bam*H1; this expression vector is referred to as the pSV3C expression plasmid in the subsequent text. The pSV3C expression plasmid places the gene encoding SV3CP under the control of an indirectly IPTG inducible T7 promoter (a modified bacterial strain i.e. of the *BL(DE3)* family, carrying the T7 RNA polymerase gene under control of the IPTG inducible *lac* operator should be used for recombinant protein expression when using pET based vectors). Carrying the Amp_r gene encoding *β*-lactamase the pSV3C expression plasmid allows for selection of successfully transformed bacterial cells by ampicillin supplemented growth media.

Proof of the correct pSV3C construct was sought by DNA sequence analysis completed by The Sequencing Service, University of Dundee, UK. Due to current technology limitations, reliable sequence information can only be obtained up to 500 bases away from the oligonucleotide primer site therefore it was necessary to sequence the gene insert in the 3' – 5' direction as well as the 5' – 3' direction to cover the entirety of the SV3CP gene. Sequence analysis confirmed the presence of the correct gene insert.

In all instances, inclusive of selenomethionine derivative protein expression, SV3CP was over-expressed in *E. coli* strain *BL21(DE3)*p*LysS*. *BL21* derivative strain *E. coli* lack the LON; and membrane bound OmpT proteases found in B strain *E. coli*. The LON and OmpT proteases are responsible for degradation of foreign, non-bacterial protein i.e. recombinantly expressed protein; their presence is quite obviously undesirable. Further, *BL21* strain cells carrying the extra plasmid, p*LysS*, increase the ease and reliability of induction of expression of recombinant protein by preventing leaky pre-induction expression. *BL21* strain cells lacking the p*LysS* plasmid can experience some low level expression of the T7 RNA polymerase prior to induction during the bacterial cells growth phase. In the instance of recombinant expression of a protease, as here, leaky pre-induction expression can cause potentially cytotoxic effects. The p*LysS* plasmid encodes T7 lysozyme that conveniently digests T7 RNA polymerase pre-induction. T7 lysozyme is expressed at relatively low levels and therefore does not interfere with recombinant protein expression, post induction, when levels of T7 RNA polymerase are relatively high.

Standard protocol when working with *BL21*(DE3)p*LysS* cells is to grow to an optical density (OD_{600}) of 0.8 with respect to un-inoculated fresh media to ensure the cells are still in the rapid growth phase at the point of induction to maximise recombinant protein expression. Growing to an optical density higher than an OD_{600} equal to 0.8 (other *BL21* derivative cells are typically grown to an OD_{600} of 1.0 prior to induction) allows *BL21(DE3)*p*LysS* cells to enter the stationary phase of the growth curve where gene expression is reduced to house-keeping proteins only; causing recombinant protein expression to be at a minimum. In all instances the bacterial cell cultures were therefore grown to a maximum OD_{600} of 0.8 pre induction.

Standard protocol was followed for transformation of *BL21(DE3)*p*LysS* with the expression plasmid; pSV3C. Native protein expression was achieved following standard protocol for *E.coli* expression systems in Luria broth where target protein over-expression is IPTG inducible (*Figure 4.1*); cells are incubated at 37°C with moderate shaking (fermenter growths are stirred) for the purpose of aerating the growth media.

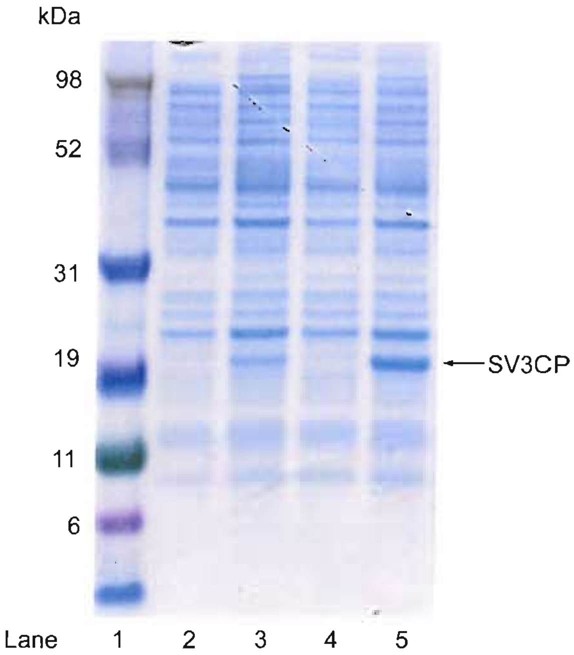


Figure 4.1: SDS PAGE analysis showing efficacy of IPTG induction of over-expression of SV3CP*. For comparison cell lysate from flask and fermenter growths are shown; Lane 1; Molecular weight markers (kDa); Lanes 2 and 3: Flask growth pre and post induction respectively; Lanes 4 and 5; Fermenter growth pre and post induction respectively. The higher protein yield from an equal volume of fermenter growth in comparison to flask growth is clearly demonstrated. In both the flask and fermenter growths post induction samples were taken 3 hours following induction of over expression with IPTG.*

To ascertain an appropriate outgrowth period, post induction, in order to maximise yield of recombinantly expressed SV3CP a time course expression experiment was completed. To these ends 10 ml of growth culture was withdrawn from a single 800 ml growth at time points; 1.0 hours, 1.5 hours, 2 hours, 2.5 hours, 3.0 hours, 3.5 hours and 4 hours post-induction (*Figure 4.2*). The crude cell lysate from each 10ml culture was analysed by SDS-PAGE to determine the optimal out-growth period. A large band corresponding to a molecular weight of approximately 19 kDa could clearly be seen in all samples, confirming expression of SV3CP, with the largest band at 3 hours post induction. Accordingly an outgrowth period post induction of 3 hours was adopted for expression of recombinant protein.

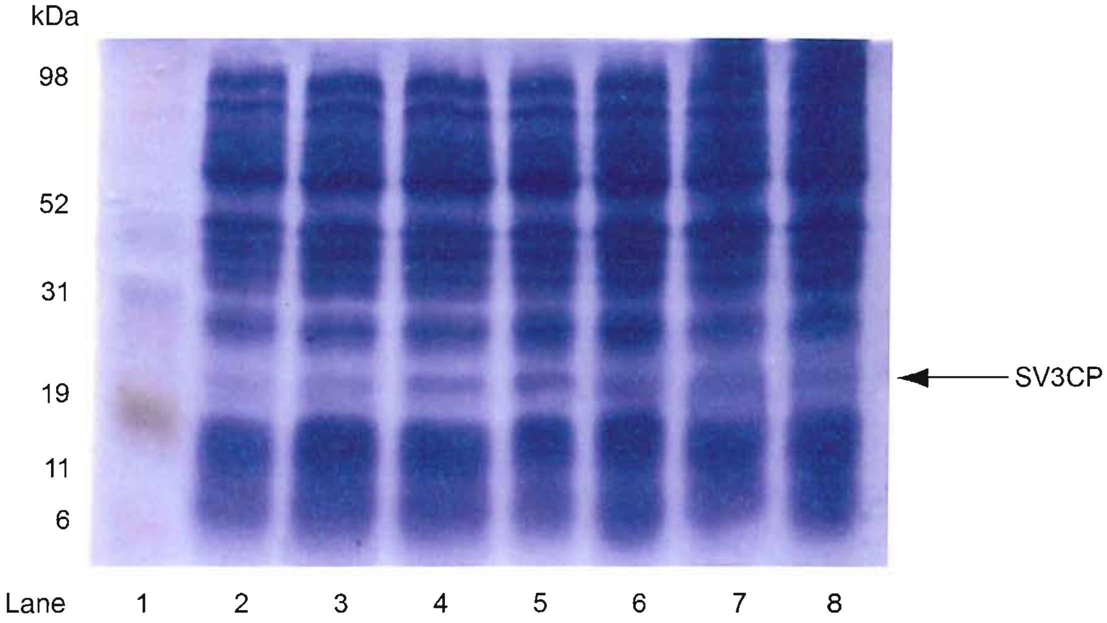


Figure 4.2: SDS PAGE analysis showing time course of increasing out growth periods following induction of over expression of SV3CP*. Lane 1; Molecular weight markers (kDa); Lane 2: 1 hour outgrowth; Lane 3: 1.5 hours outgrowth; Lane 4: 2 hours outgrowth; Lane 5: 2.5 hours outgrowth; Lane 6: 3 hours outgrowth; Lane 7: 3.5 hours outgrowth; Lane 8: 4 hours outgrowth.*

Preparative recombinant SV3CP expression was routinely completed on a large scale by fermentation in 10 L quantities. Such was the success of this approach, typical yields would be 20 mg/L of growth media following purification compared to typical yields of 15 mg/L by flask growth.

4.1.2 Expression of a selenomethionine derivative SV3CP

Expression of a selenomethionine derivative SV3CP was achieved by adaptation of a protocol reported by Ramakrishnan (162) where the need for a traditional methionine auxotroph *E. coli* strain is dispensed with and expression is completed in the same bacterial strain as used for native recombinant protein expression. The protocol followed is the same as for native SV3CP expression, up to point of induction with IPTG. At this point the cell culture is transferred to a M9 minimal media by centrifugation of the LB growth media to pellet cells, then their re-suspension in M9 media. M9 media lacks all amino acids and acts solely as a carbon and nitrogen source. Cells were then grown in this basic media for 30 minutes, the rationale being that all endogenous methionine would be used in expression of bacterial proteins during this period. Following this, a synthetic cocktail of all amino acids was added to a final concentration of 40 µg/ml of growth media with the exception of methionine. Selenomethionine was then added to a final concentration of 40 µg/ml of growth media; a relatively high concentration and sufficient to inhibit endogenous methionine synthesis. Hence all protein expressed from this point onwards contain the exogenous source of selenomethionine. In addition to amino acids a mix of essential vitamins and minerals was added to the growth media, the mix contains; riboflavin, niacinamide, pyridoxine monohydrochloride, and thiamine. A similar time course experiment to that for native protein expression was completed to determine an optimal out growth period to maximise yield of the selenomethionine derivative SV3CP. Due to the cyto-toxic property of selenomethionine a suitable outgrowth period is even more critical than with native protein expression. Indeed, when grown for longer than 3.5 hours post induction cells were seen to die as reflected in a reduction in OD_{600} of the culture and of relative quantities of SV3CP, as judged by SDS-PAGE. The time course experiment indicated an optimal out-growth period of 2.5 hours.

Due to the expense of selenomethionine and the need for relatively smaller amount of selenomethionine derivative SV3CP compared to native, selenomethionine growths were completed in 1 L baffled flasks containing 400 ml of growth media only. A typical yield of selenomethionine derivative SV3CP by flask growth was typically 12 mg/L of growth media following purification.

4.1.3 Purification of SV3CP to homogeneity by ion exchange chromatography

An identical method of purification was adopted for both native and selenomethionine derivative SV3CP. The SWISSPROT PROTEINPREDICT tool estimated a *pI* (isoelectirc point) of 8.1 for SV3CP. Following standard protocol a starting pH of 7.45 (between 0.5 and 1.0 pH units below the isoelectric point), was selected for the buffers used during purification by cation exchange chromatography. Frozen cell pellets from 5 L worth of growth were thawed and resuspended in 10 mM potassium phosphate buffer, pH 7.45 with 5 mM β-mercaptoethanol (buffer A) added to maintain reducing conditions. The cell suspension was then homogenised manually before cells were lysed by sonication. Cell debris was pelleted by centrifugation, the supernatant was then filtered by use of a 0.45 μm syringe filter. Protein purification was completed over two columns; column 1: SP Sepharose (Fast Flow grade); column 2: SOURCE 15S, both at room temperature. The filtered crude cell lysate was loaded directly onto the SP Sepharose column (column size: 16 x 200 mm). Loading crude, if filtered, cell lysate directly onto an ion exchange column without a broad spectrum purification step such as ammonium sulphate protein precipitation may seem a heavy handed approach; however it seemed truer to the protein purification maxim of "as few purification steps as possible" and produced good results. Prior to protein elution, by running a salt gradient, proteins not bound at pH 7.45 were eluted by running buffer A through the column until a steady baseline absorbance (measured at 280 nm) was achieved. Bound protein was then eluted by increasing the strength of the ionic environment by running a salt gradient from 0 M to 1 M NaCl. During early purification attempts 5 ml protein fractions were collected and analysed by SDS PAGE to identify the fraction containing SV3CP. SV3CP was seen to elute over a single large peak of the 280 nm absorption chromatograph at approximately 200 mM NaCl (only an approximate value can be offered since the FPLC used did not possess a conductivity meter) with only slight shoulders either side of the peak, indicating relatively pure protein, indeed this was proven by SDS PAGE (*Figure 4.3*). Such was the size of the peak relative to other protein peaks, it became quite easy to identify the fraction containing SV3CP in subsequent purifications upon the 280 nm chromatograph alone, therefore routine analysis of protein fractions by SDS PAGE was dispensed with once a reliable purification protocol had been established.

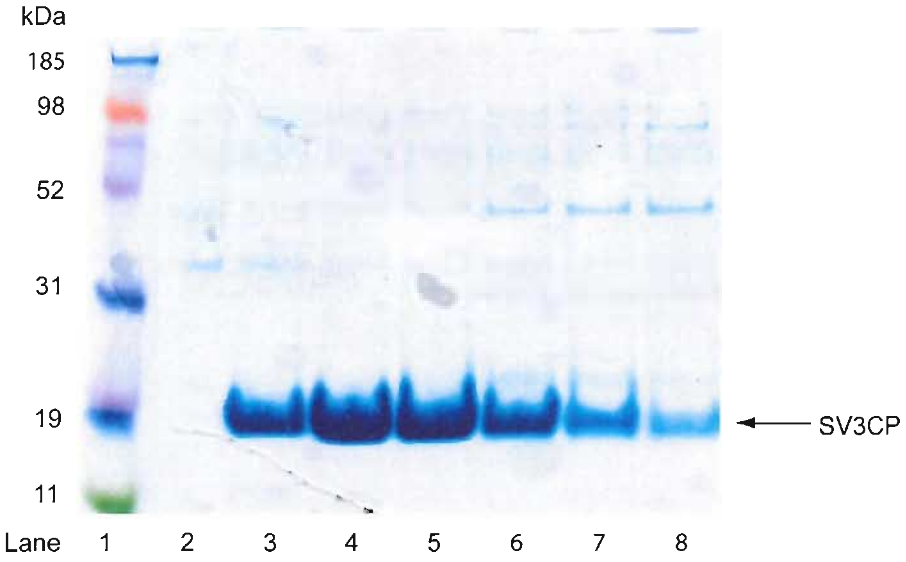


Figure 4.3: SDS PAGE analysis showing efficacy of SV3CP purification by SP Sepharose cation exchange chromatography*. Lane 1: Molecular weight markers (kDa); Lane 2; fraction before SV3CP peak on 280nm chromatograph, notable due to the complete lack of SV3CP indicative of a sharp protein peak; Lanes 3 – 8 inclusive; consecutive fractions comprising the entire SV3CP peak. In all lanes, bands in addition to those at approximately 19 kDa can be seen and are indicative of contaminating bacterial proteins.*

Subsequent to SP Sepharose chromatography fractions containing SV3CP were desalted using a G25 Sephadex (Fine grade; column dimensions: 26 x 400 mm) gel filtration column. The entire desalted sample was purified in a single attempt by SOURCE 15S chromatography (column dimensions: 10 x 200 mm). Protein was eluted by salt gradient 0 M – 1 M NaCl, with a single very sharp large peak on the 280nm absorption chromatograph at approximately 400 mM NaCl. Prior to storage or immediate use, the protein was desalted by gel filtration (G25 Sephadex). For long term storage, glycerol was added to the SV3CP solution to a final concentration of 50%, the solution was then flash frozen, drop-wise in liquid nitrogen, and stored at -80°C. SV3CP stored in this fashion showed full activity even following a year of storage. When needed, frozen SV3CP was thawed and glycerol removed by gel filtration (G25 Sephadex) and notably even proved suitable for use in crystallisation trials where entirely fresh protein is normally sought. Analysis by SDS PAGE (*Figure 4.4*) of the fully purified protein revealed a highly pure SV3CP sample, confirmed by electro-spray ionisation orthogonal acceleration time of flight mass spectrometry (ESI-oa-TOF-MS) (*Figure 4.5*).

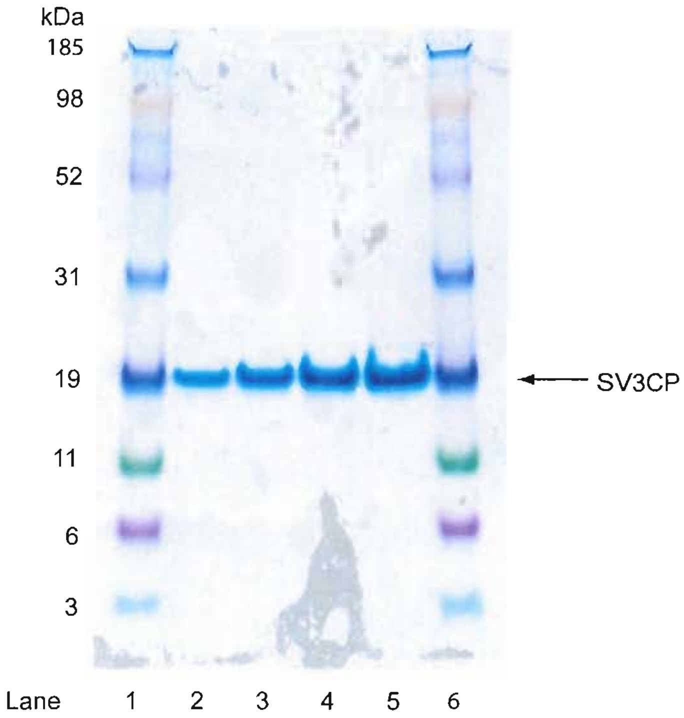


Figure 4.4: SDS PAGE analysis showing efficacy of SV3CP purification by SOURCE 15S cation exchange chromatography*. Lanes 1 and 6: Molecular weight markers (kDa). Lanes 2 – 4 inclusive; Repeats, using increasing sample size on gel, of the single fraction comprising the entire SV3CP FPLC 280 nm chromatograph peak. A lack of any visible bands in addition to that corresponding to protein of approximately 19 kDa is indicative of a highly pure protein sample.*

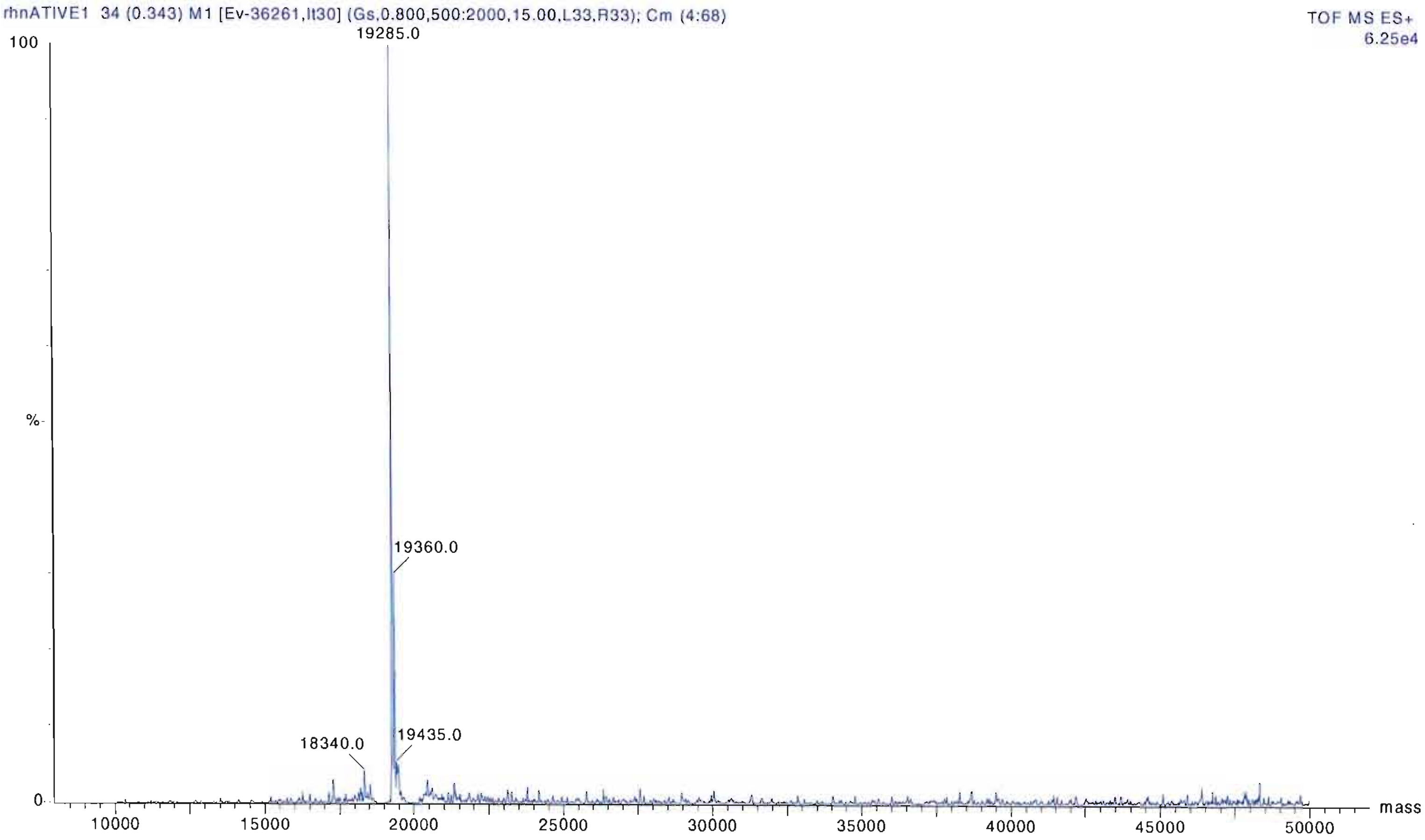


***Figure 4.5: ESI-oa-TOF mass spectra of fully purified native SV3CP**. A lack of any major peaks at masses other than that corresponding to a single molecule of SV3CP; 19, 285 kDa is indicative of a highly pure protein sample and therefore an effective purification procedure.*

4.1.4 Confirmation of a selenomethionine derivative SV3CP

Comparison of the single main peaks of the ESI-oa-TOF spectra of native SV3CP and the selenomethionine derivative (*Figure 4.6*) reveals a mass difference of 235 Da, equivalent to the exact mass difference resulting from the substitution of 5 methionines for selenomethionines. Each molecule of SV3CP possesses 5 methionine residues; this result was therefore indicative of a 100% incorporation of selenomethionine in place of methionine in all molecules of the selenomethionine derivative SV3CP.

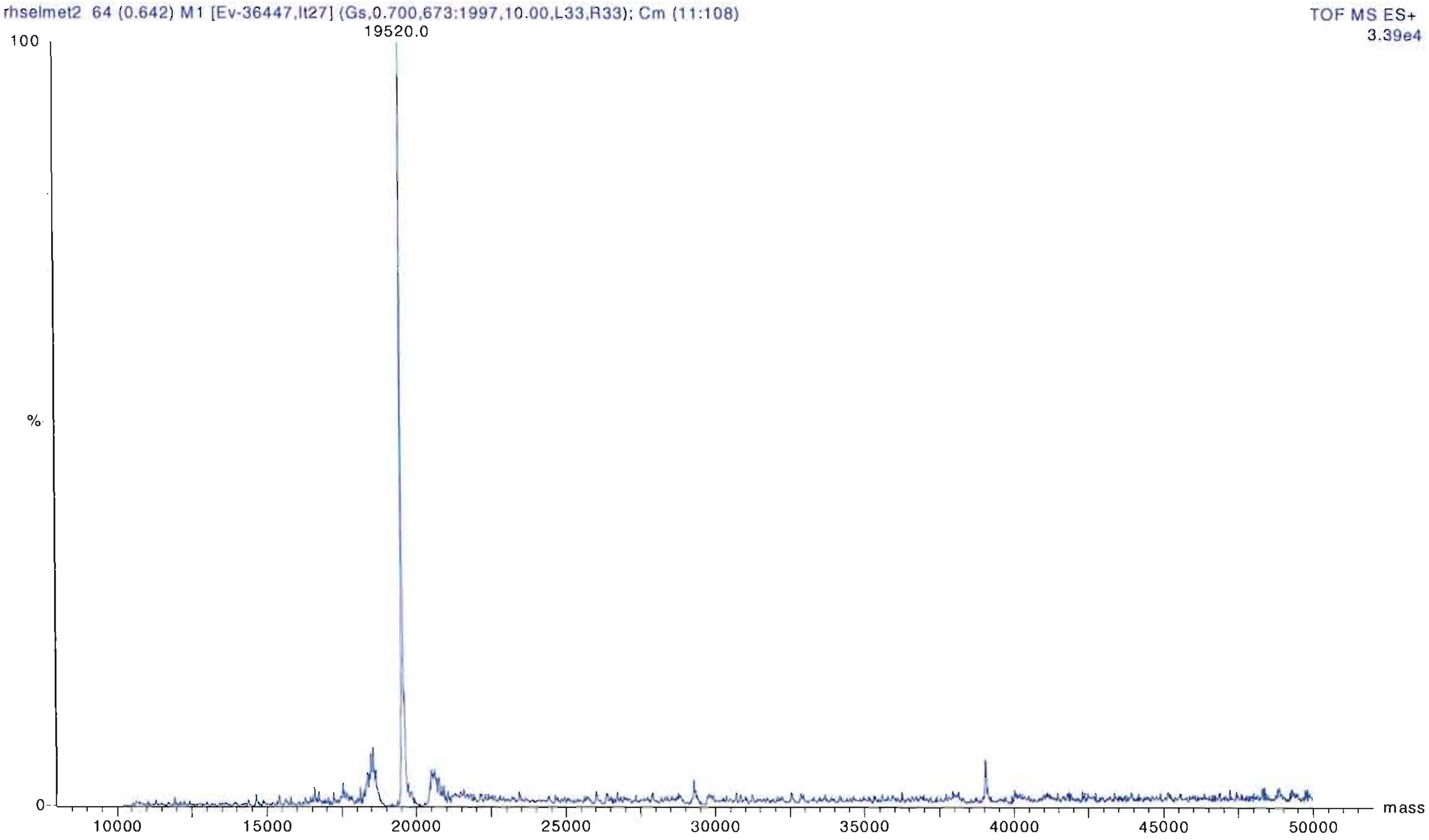


***Figure 4.6: ESI-oa-TOF mass spectra of selenomethionine derivative SV3CP**. The single main peak at 19, 520 kDa corresponds to a single molecule of SV3CP with an increased mass over native SV3CP equivalent to the exact calculated mass for 100% incorporation of selenomethionine.*

4.1.5 Typical SV3CP yield following purification

Protein concentration was determined following each purification step by either BIORAD protein assay (based on the Bradford method of determining protein concentration) or spectrophotometrically by absorption at 280nm where protein concentration is defined by:

$C = A / El$

Where A equals the absorption at 280 nm, E is the extinction co-efficient of the protein; of 1.317 for SV3CP, and l is the path length. The extinction co-efficient is easily calculated adopting the following equation:

$E = (5{,}700 \times nW) + (1{,}300 \times nY) / Mr$

Where nW equals the number of tryptophans and nY the number of tyrosines.

Protein yields at each stage of purification are shown in Table 4.1 and are reflective of 5L worth of a fermenter growth.

Step	Protein concentration by BIORAD mg/ml	Volume of sample /ml	Total protein yield /mg
SP Sepharose	4.17	30	125.1
G25 Sephadex (desalting step)	2.13	53	113.2
Source 15S	9.77	10	97.7
G25 Sephadex (desalting step)	5.78	26	92.3

Table 4.1: Typical SV3CP purification table*. Values shown refer to purification of 5 L of fermenter growth.*

4.1.6 Qualitative assay of SV3CP activity

Activity of fully purified SV3CP was qualitatively assessed by incubation of 5 µl of SV3CP in solution (10 mM phosphate buffer pH 7.45, 5 mM *β*-mercaptoethanol; typically 5 mg/ml after final desalting) with 5 µl of a 100 mg/ml solution of the chromogenic peptide Ac-EFQLQ-pNA (see *Section 3.2*) in DMSO. Active SV3CP caused a strong yellow colour to develop after 2 -3 minutes incubation at room temperature due to cleavage of the peptide's *C*-terminal chromophore. Activity assays were routinely completed on this small scale and were intended as a qualitative assessment of SV3CP activity only. For reasons of economy (to reduce usage of the precious chromogenic peptide), it was not possible to routinely assess activity of SV3CP quantitatively for every purification completed. For more rigorous kinetic assay see *Section 4.2.4*.

4.2 Synthesis of a series of chromogenic peptides

4.2.1 Successful synthesis of a series of chromogenic peptides

Following the protocol detailed in *Section 3.2* the chromogenic peptides: Ac-QLQ-*p*NA, Ac-FQLQ-*p*NA, Ac-EFQLQ-*p*NA and Ac-DEFQLQ-*p*NA were successfully synthesised. Matrix assisted laser desorption ionisation quadrapole time of flight mass spectroscopy (MALDI-Q-TOF-MS) prior to purification produced spectra clearly showing major peaks corresponding to the calculated mass of each peptide (spectra not shown).

4.2.2 Purification of the chromogenic peptides by reverse phase chromatography

Purification of the chromogenic peptides proved of far greater difficulty than their synthesis since all displayed a distinct insolubility in polar solvent. Some hydrophobic nature had been expected as residues phenylalanine and leucine can be considered highly hydrophobic, however, the degree of hydrophobicity remained a surprise.

For all peptides full dissolution could only be achieved in a solution of organic solvent content of 60% minimum i.e. 60% acetonitrile in water. This is problematic for purification by HPLC. At such high organic solvent concentrations most reverse phase columns will not interact strongly enough with the mobile phase to allow retention of any species sufficiently to allow acceptable separation by further increase in eluent hydrophobicity. Further, if the peptide samples had been loaded onto a column in a 60% acetonitrile solution, to achieve any species separation the peptide would have to possess a higher affinity for the stationary phase than the mobile phase i.e. bind to the column matrix. A reverse phase column that would bind any of the peptides when dissolved in a solution of 60% acetonitrile could not be identified (later it was found that between 40% and 45% acetonitrile in water was sufficient to elute all peptides from a C18 reverse phase column). Therefore it was necessary to identify an alternative organic solvent to acetonitrile in which to dissolve the crude peptides in preparation for purification. To these ends 10 mg of crude peptide (irrespective of peptide) was fully dissolved in 100 µl of DMSO. Since DMSO will not bind to the C18 column matrix but rapidly elute, a risk existed that the peptide would precipitate at the top of the column as the DMSO component separates from the sample. Precipitated peptide would be irretrievable, and also damage the column. To reduce this risk, the peptides sample was made slightly polar by addition of 100 µl of water; although this had the undesirable effect of producing a turbid sample. The sample was then loaded in its entirety onto the C18 reverse phase column and an elution gradient of increasing hydrophobicity was run over 80 ml at a flow rate of 4 ml/min from 30% to 50% acetonitrile (+0.1% TFA) in water. Though not ideal to apply a turbid sample to a high grade reverse phase column, or indeed inject a turbid sample into a HPLC system at all, in this instance it remained the most convenient and successful approach. The same purification was applied to and successful for all four peptides. Some variation was seen in the concentration of acetonitrile at which each peptide eluted, though it was always between 40% and 45% acetonitrile. Absorbance chromatographs for all samples showed good baseline separation with a single easily identifiable peak corresponding to pure product. In conjunction with the absorbance chromatographs purification success was judged by MALDI-Q-TOF-MS. Individual spectra pertaining to each peptide clearly showed single major peaks corresponding to the respective calculated mass indicating highly pure samples (*Figures 4.7 to 4.10*). On a technical issue; samples analysed by MALDI-Q-TOF-MS were used directly as they eluted from the C18 reverse phase column i.e. no further preparative steps were required before analysis.

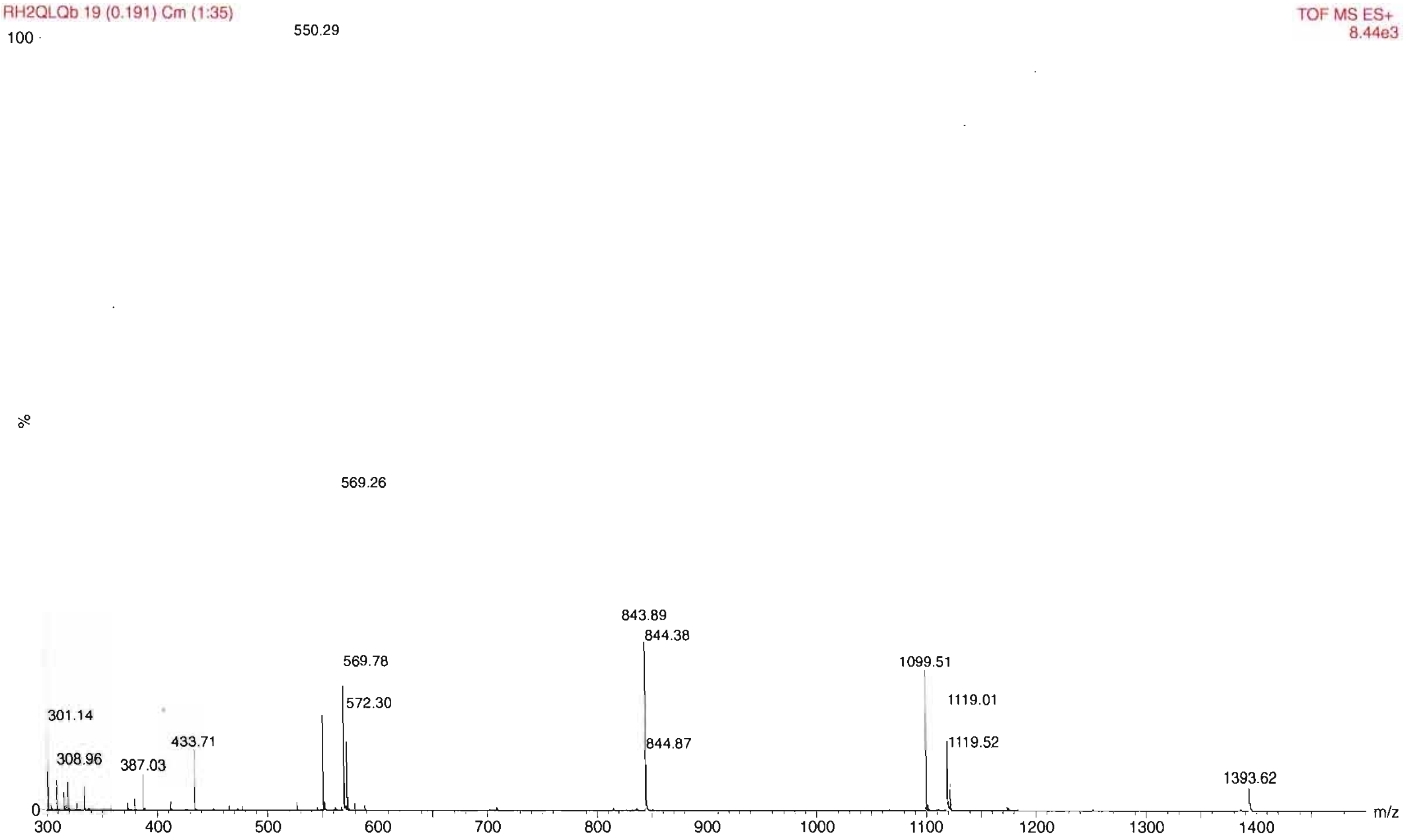


Figure 4.7: MALDI-Q-TOF mass spectrum of purified chromogenic peptide Ac-QLQ-pNA*. Main species peak seen at exact calculated weight of Ac-QLQ-pNA at 550 Da; a lack of any other significant peaks is indicative of a pure sample and therefore efficacy of purification procedure.*

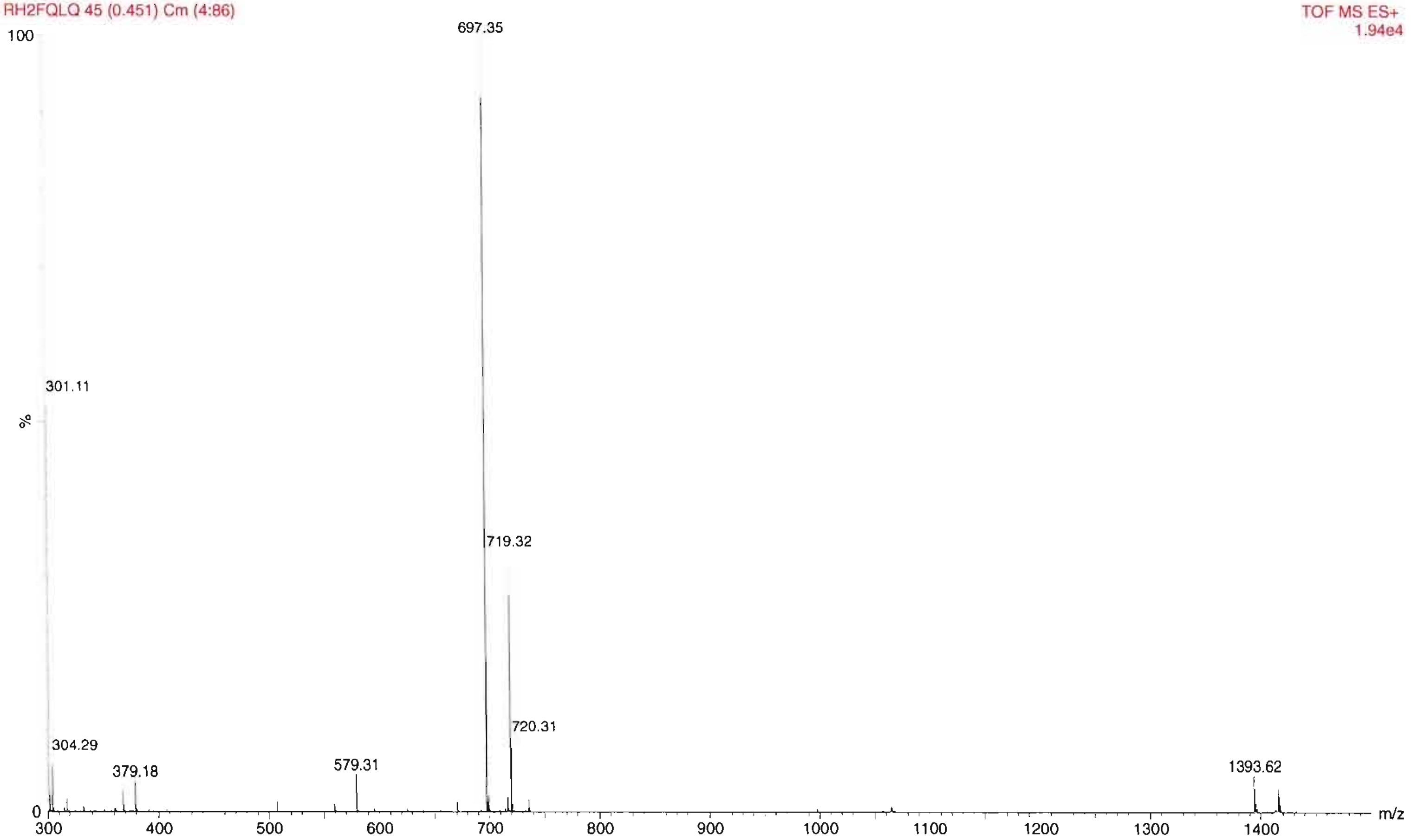


Figure 4.8: MALDI-Q-TOF mass spectrum of purified chromogenic peptide Ac-FQLQ-pNA. *Main species peak seen at exact calculated weight of Ac-FQLQ-pNA at 697 Da; a lack of any other significant peaks is indicative of a pure sample and therefore efficacy of purification procedure.*

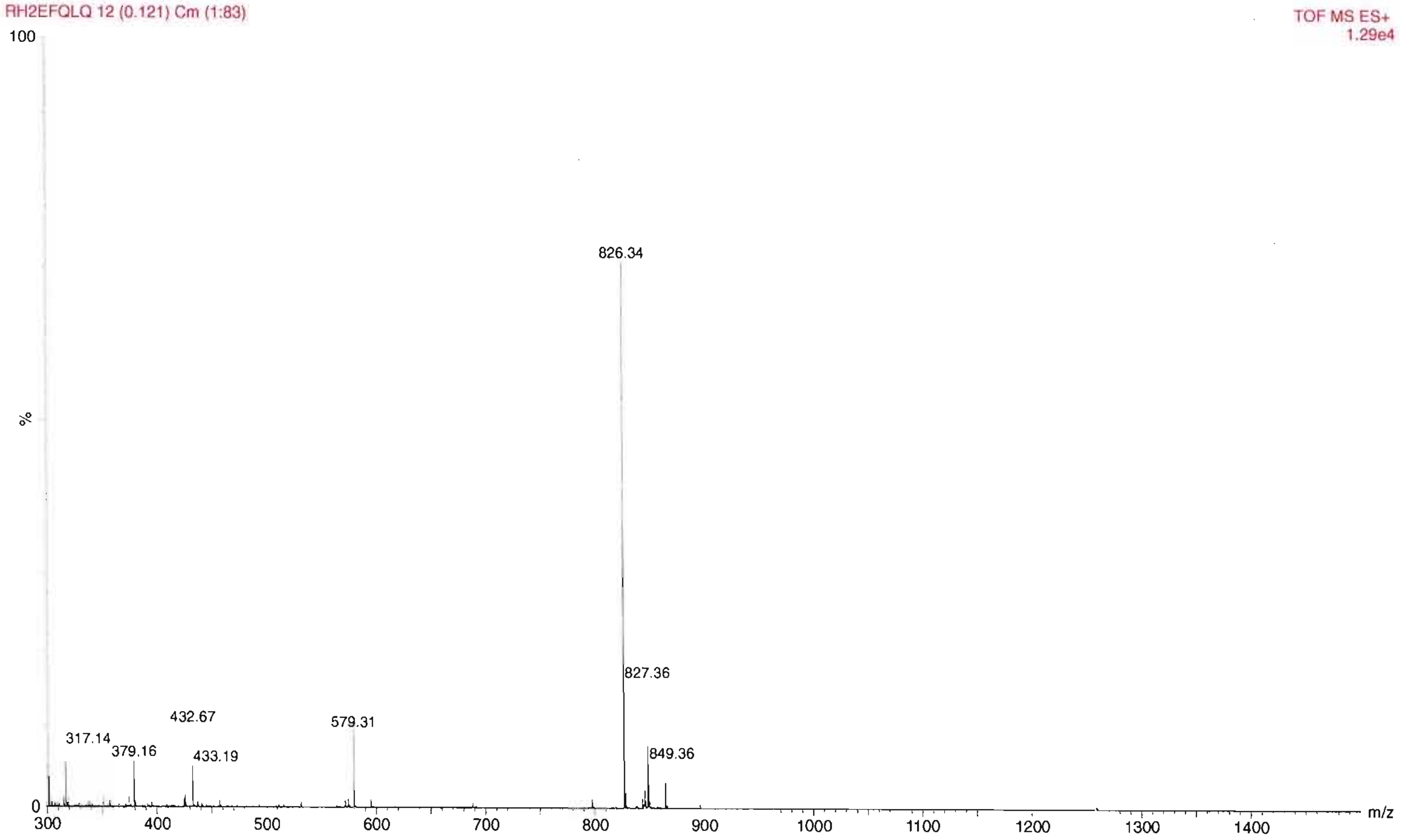


***Figure 4.9: MALDI-Q-TOF mass spectrum of purified chromogenic peptide Ac-EFQLQ-pNA**. Main species peak seen at exact calculated weight of Ac-EFQLQ-pNA at 826 Da; a lack of any other significant peaks is indicative of a pure sample and therefore efficacy of purification procedure.*

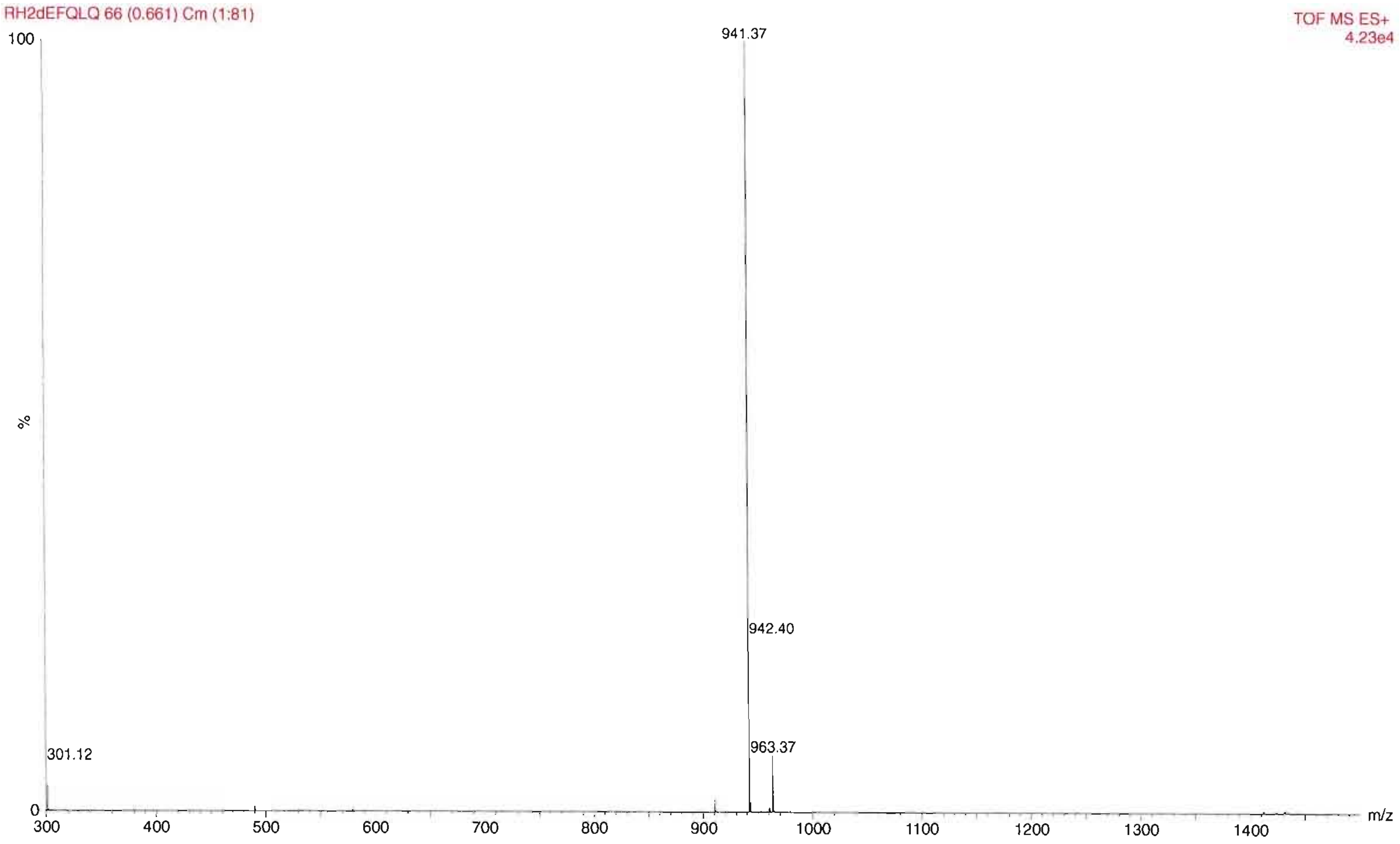


Figure 4.10: MALDI-Q-TOF mass spectrum of purified chromogenic peptide Ac-DEFQLQ-pNA. *Main species peak seen at exact calculated weight of Ac-DEFQLQ-pNA at 941 Da; a lack of any other significant peaks is indicative of a pure sample and therefore efficacy of purification procedure.*

4.2.3 Typical chromogenic peptide product yield following purification

Most probably due to the rather rudimentary approach taken in purification comparison of total masses pre- and post-purification showed an approximate yield of only 60% for all peptides. A 40% product loss could not be attributed solely to the removal of impurities during purification since the reverse-phase chromatography 216 nm absorption chromatograph and MALDI-Q-TOF spectra prior to purification show relatively pure samples i.e. little contamination as a result of incomplete reactions during synthesis. Losses must therefore be attributed to precipitation of product on the reverse-phase column during purification.

4.2.4 Kinetic assay of the SV3CP activity by chromogenic peptides

Initial assays showed that three of the four chromogenic peptides were cleaved by SV3CP as qualitatively judged by the development of a yellow colour corresponding to liberation of the chromophore: *p*-nitroaniline.

Prior to use in kinetic assays, SV3CP was buffer exchanged by gel filtration into the following buffer: 100 mM TRIS pH 8.5, 5 mM β-mercaptoethanol to provide the optimal pH for SV3CP to achieve a maximum rate of activity. Quantitative kinetic analysis was subsequently performed using a 96 well micro-plate reader. Absorbance was measured at 480 nm over 2 minutes at 10 second intervals following addition of 290 μl of a 2 mg/ml solution of SV3CP to 10 μl of chromogenic peptide previously dissolved in 100% DMSO at concentrations: 5 mM, 10 mM, 15 mM, 20 mM, 25 mM. All assays, for all four peptides, were completed in triplicate.

Measurement of initial reaction velocity V_0 revealed peptides; Ac-DEFQLQ-*p*NA and Ac-FQLQ-*p*NA to show a similar rate of cleavage to each other and approximately 50% less than Ac-EFQLQ-*p*NA. Ac-QLQ-*p*NA showed an extremely low V_0 (barely above the baseline reading) corresponding to a low rate of turnover by SV3CP (*Figure 4.11*). Since Ac-EFQLQ-*p*NA displayed the greatest initial rate of cleavage it subsequently was used in routine assay of a SV3CP activity in preference to the other peptides. Whilst these results remain of un-publishable quality, the X-ray structure of SV3CP in complex with a peptidyl inhibitor later showed that the SV3CP active site cleft 'S' sites to co-ordinate with a maximum of 5 peptide residues i.e. the same number of residues as the peptide portion of Ac-EFQLQ-pNA, providing further evidence for its preferential rate of cleavage.

The data represents only very basic kinetic analysis that show the relative rates at which each of the chromogenic peptides were cleaved. More rigorous kinetic analysis had been planned though time constraints prevented these from being completed. These preliminary results did however offer sufficient information on which to base the peptidyl inhibitor (see *Section 4.3*).

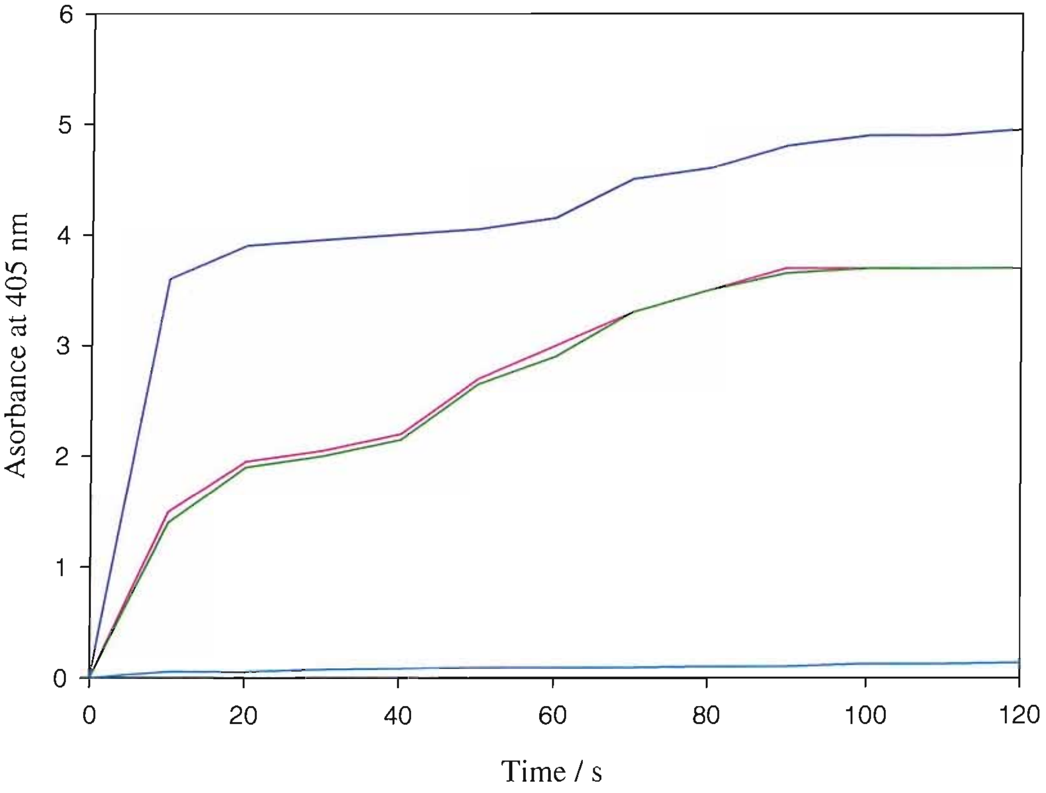


Figure 4.11: Rudimentary comparative kinetic analysis of the four chromogenic peptides. *Ac-QLQ-pNA (cyan), Ac-FQLQ-pNA (green), Ac-EFQLQ-pNA (blue) and Ac-DEFQLQ-pNA (pink). Ac-EFQLQ-pNA experiences the greatest initial rate of cleavage, V_0.*

4.3 Synthesis of a highly potent inhibitor of SV3CP activity

4.3.1 Successful synthesis of a Michael acceptor peptidyl inhibitor

Following the protocol described in *Sections 3.5 - 3.11*, the Michael acceptor peptidyl inhibitor (MAPI); Ac-EFQLQ-propenyl-ethyl-ester was successfully synthesised. MALDI-Q-TOF-MS was used after each stage in synthesis to ascertain production of the desired product. The mass spectra included in *Figures 4.12* to *4.16* provide evidence of successful synthesis of MAPI.

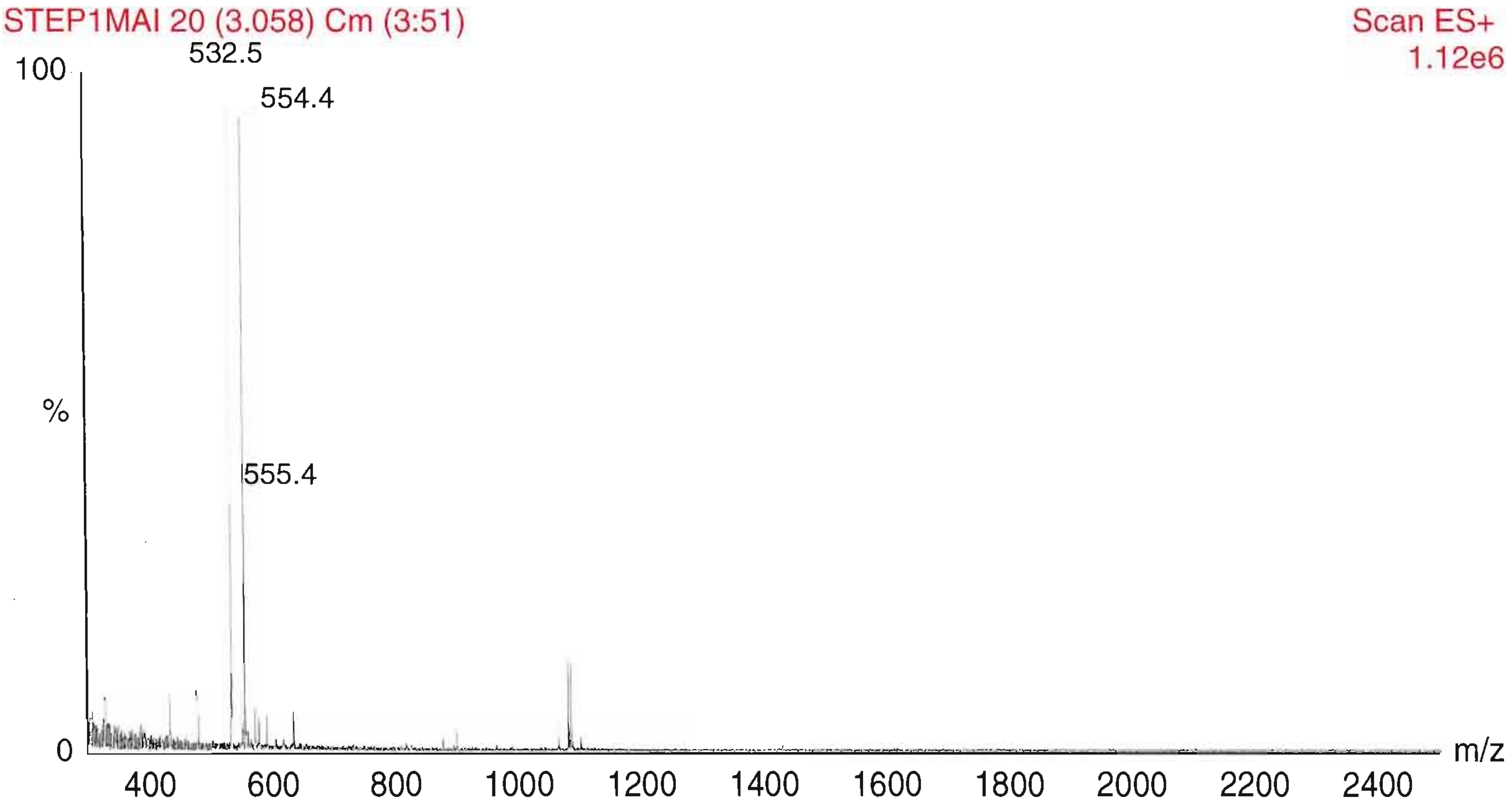


Figure 4.12: Synthesis of MAPI: step 1: Generation of the Weinreb amide*; MALDI-Q-TOF mass spectrum of crude Boc-L-Tr-Gln-$N(OCH_3)CH_3$; main species peak seen at exact calculated weight of 532.5 Da.*

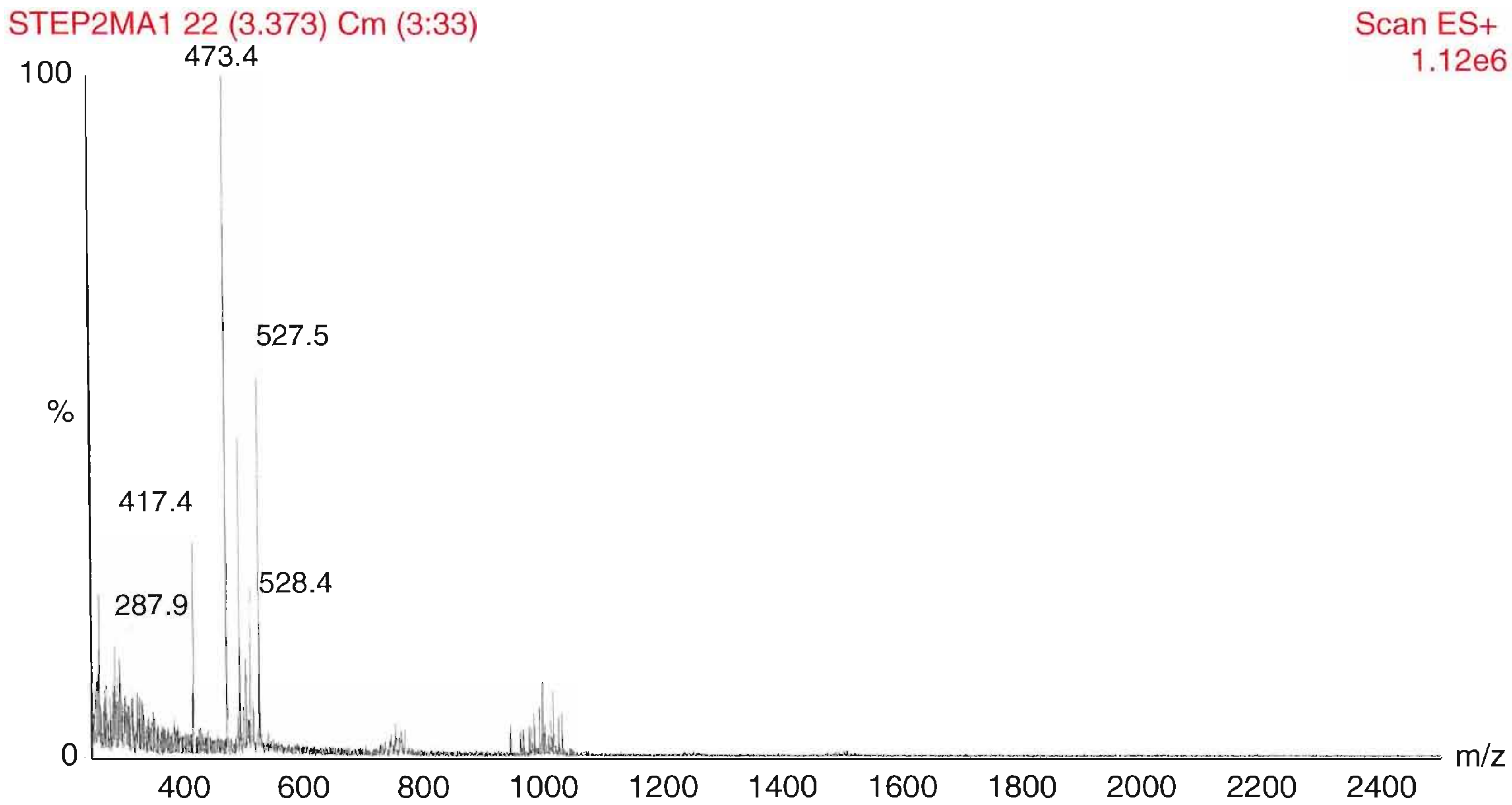


Figure 4.13: Synthesis of MAPI: step 2: Reduction to the aldehyde*. MALDI-Q-TOF mass spectrum of crude Boc-L-Tr-Gln-H; main species peak seen at exact calculated weight of 473.4 Da.*

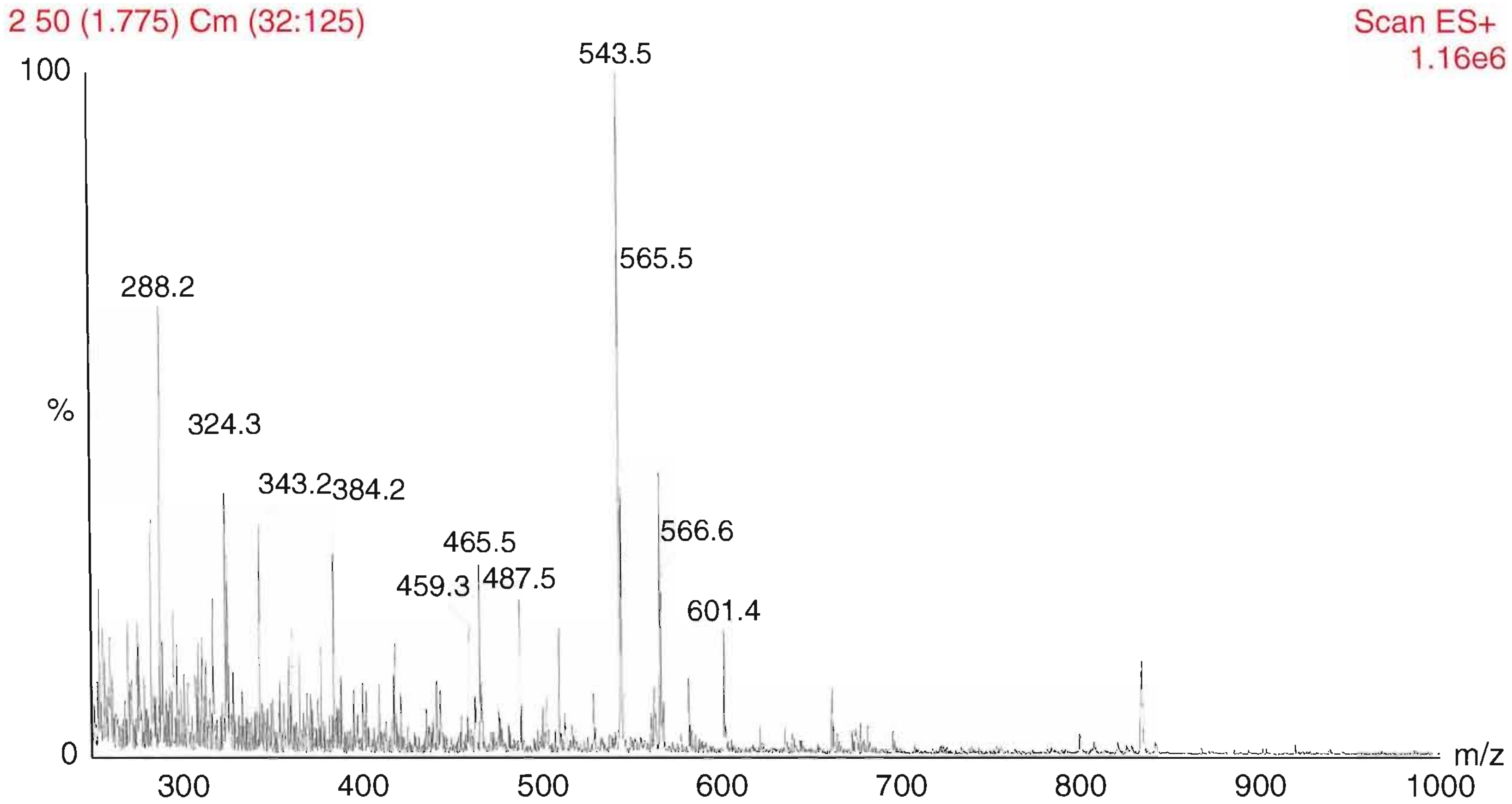


***Figure 4.14: Synthesis of MAPI: step 3: The Wittig-Horner Reaction**. MALDI-Q-TOF mass spectrum of crude Boc-L-Tr-Gln-propenyl-ethyl-ester; main species peak seen at exact calculated weight of 543.5 Da.*

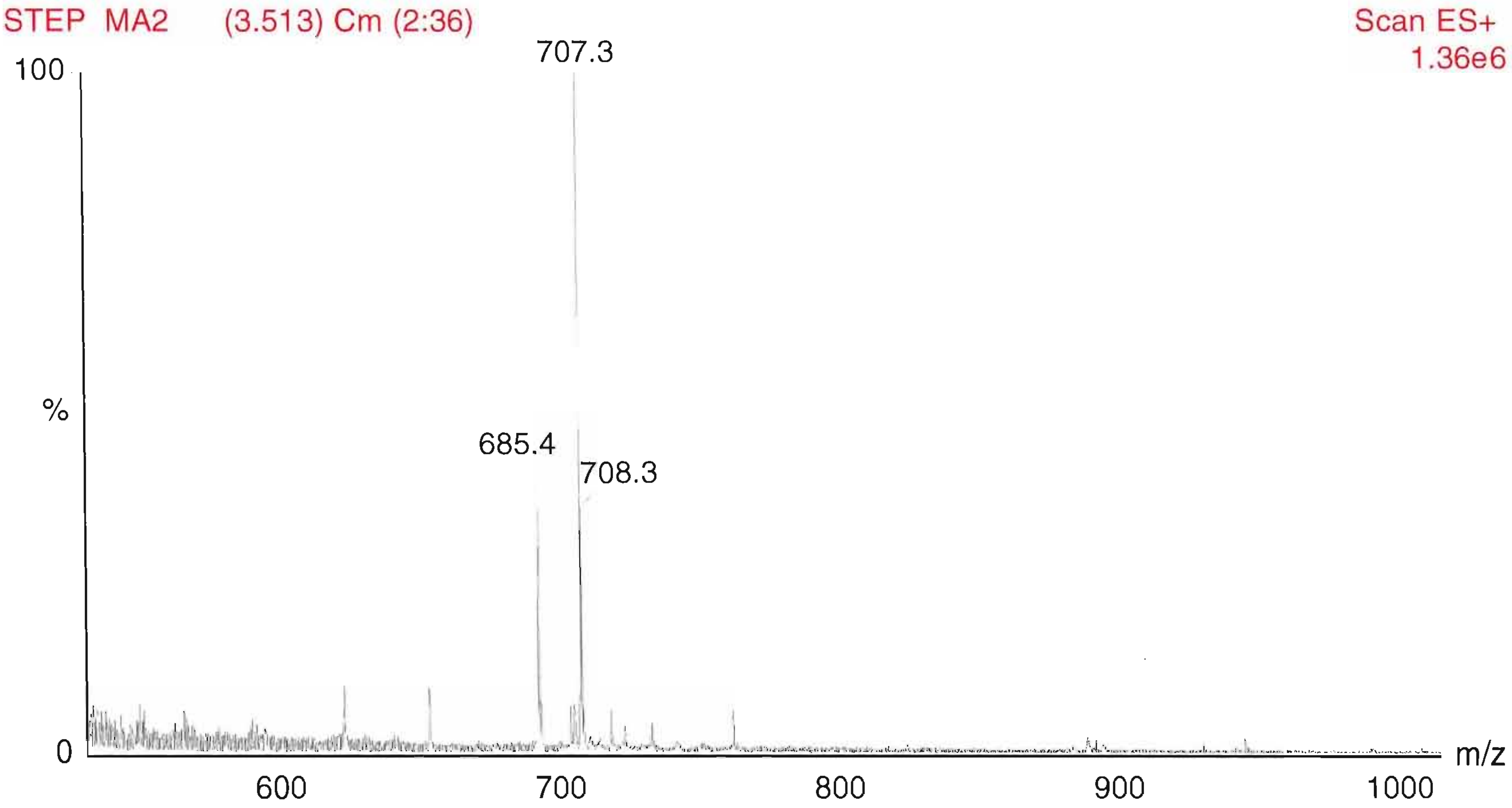


Figure 4.15: Synthesis of MAPI: step 4: Synthesis of the peptide portion. *MALDI-Q-TOF mass spectrum of crude Ac-EFQL; main species peak seen at exact calculated weight of 707.3 Da.*

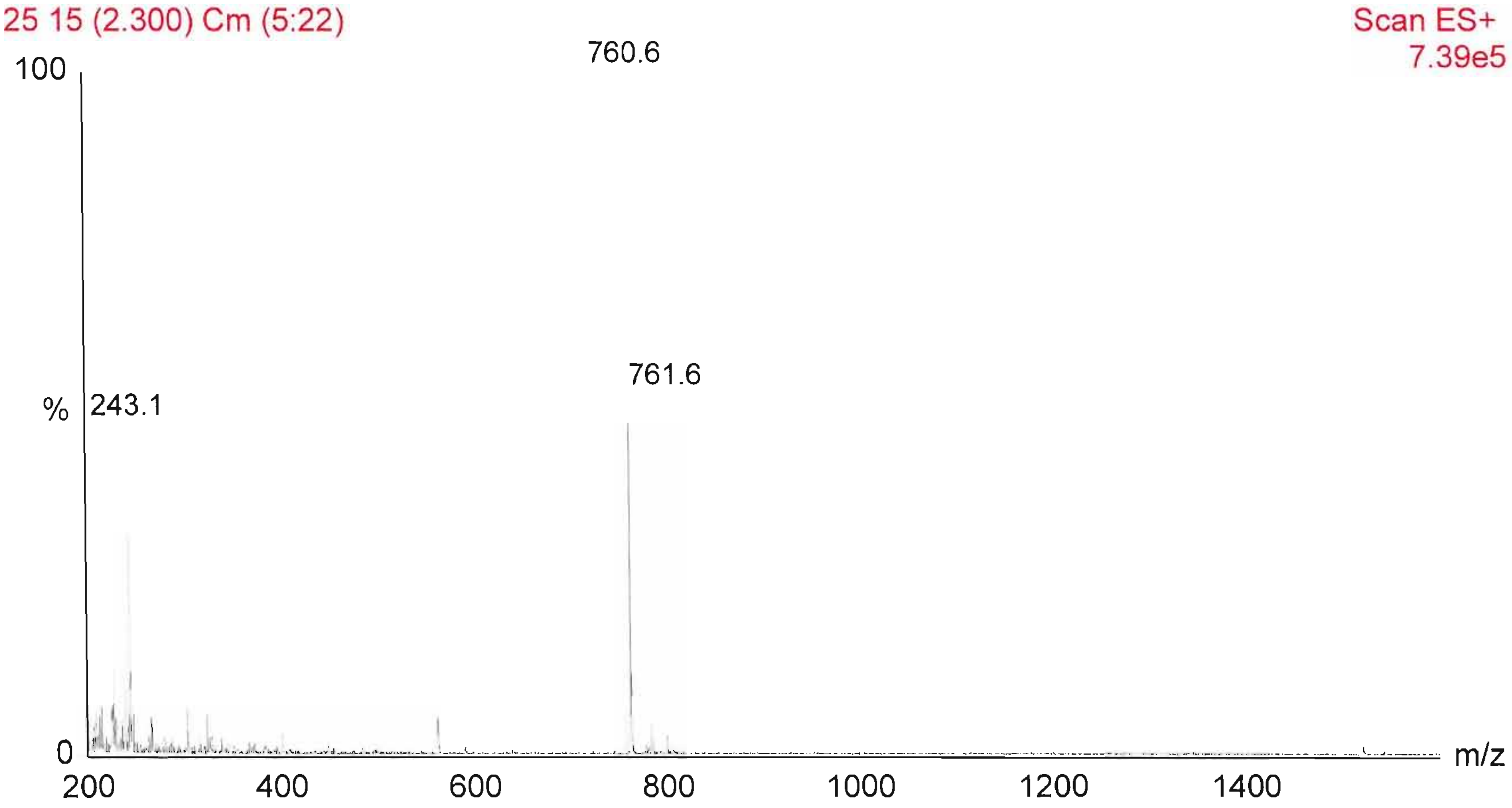


Figure 4.16: Synthesis of MAPI: step 5: Coupling of Boc-L-Tr-Gln-propenyl-ethyl-ester to Ac-EFQL and cleavage of protecting groups. *MALDI-Q-TOF mass spectrum of crude Ac-EFQLQ-propenyl-ethyl-ester; main species peak seen at exact calculated weight of 760 Da.*

4.3.2 Purification of MAPI by reverse phase chromatography

Issues of solubility, as found with the chromogenic peptides were encountered during purification of the final product: Ac-EFQLQ-propenyl-ethyl-ester. A similar approach to purification was therefore employed; 10 mg of crude MAPI was completely dissolved in 100 µl DMSO; the solution was then made to turn turbid (as observed with the chromogenic peptides) by addition of 100 µl of analytical grade water before loading directly onto a C18 reverse-phase column. A hydrophobic elution gradient was then run from 20% to 60% acetonitrile (+0.1% TFA) in water with a single major peak seen at 48% acetonitrile. Subsequent analysis by MALDI-Q-TOF-MS (*Figure 4.17*) of the reverse-phase chromatography 216 nm absorption chromatograph major peak revealed a species matching exactly the calculated mass: 760 Da, of the final MAPI product. With the absence of any other large peaks the spectra indicated a highly pure sample.

Evidence of the MAPI as highly specific and potent inhibitor of the SV3CP is offered in *Section 4.4.9.3* of this chapter.

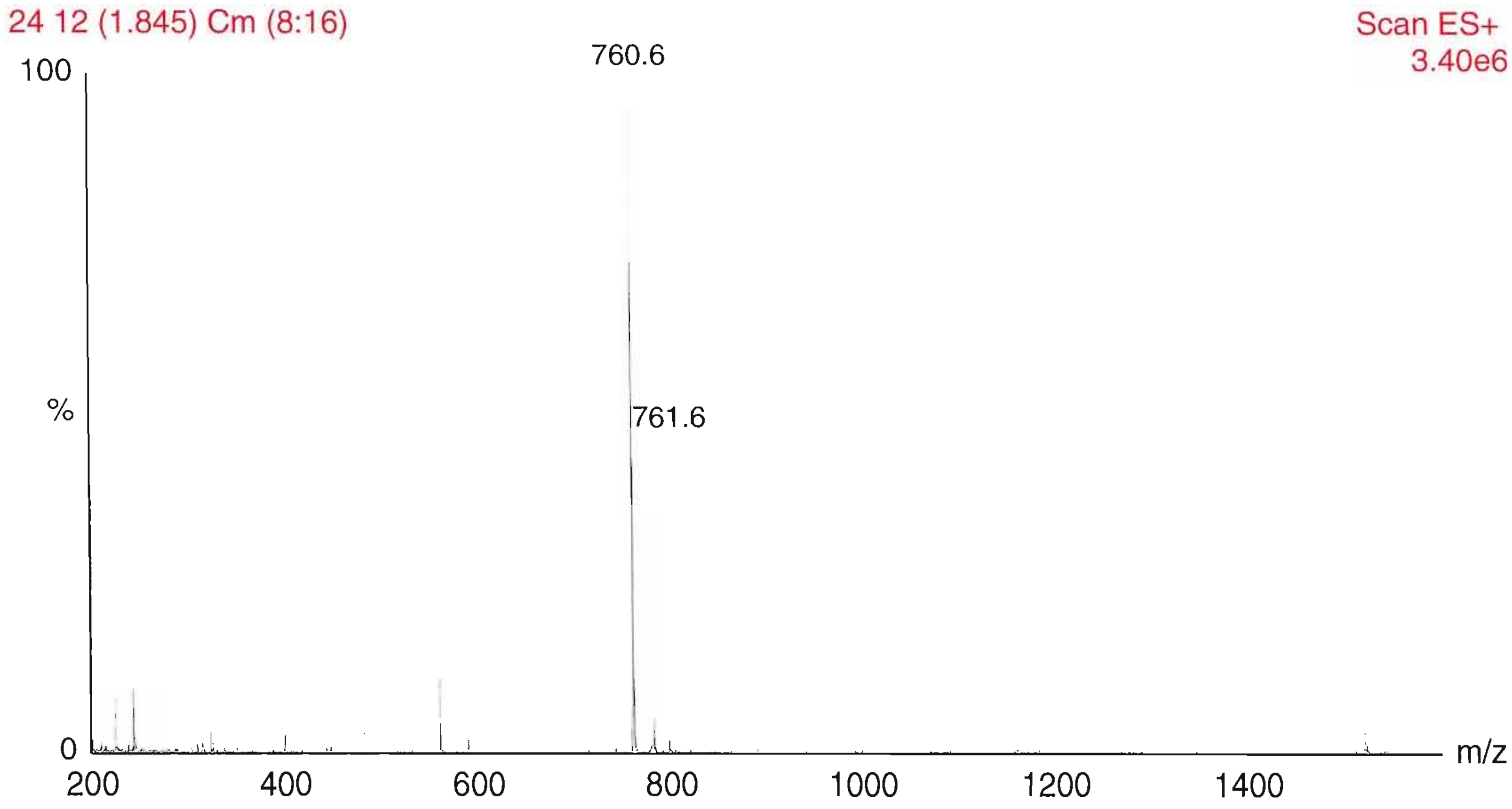


Figure 4.17: MALDI-Q-TOF mass spectrum of the purified Ac-EFQLQ-propenyl-ethyl-ester (MAPI)*. showing a main species peak at the exact calculated weight of 760 Da. Note the absence of peaks corresponding to reaction by-products as seen in the crude sample; indicative of a efficient purification procedure.*

4.4 Crystallisation and X-ray structural solution of SV3CP

4.4.1 Crystallisation of native SV3CP

Recombinantly expressed SV3CP was either used immediately following the final stage of purification or thawed from frozen stock and glycerol (50% v/v for storage at -80 °C) removed by G25 Sephadex gel filtration chromatography and subsequently concentrated to 3 mg/ml by use of 10 kDa molecular weight cut off Centricon concentration vessels. Throughout gel filtration to remove glycerol (frozen stock only) and concentration steps the SV3CP remained in a 10 mM phosphate buffer at a pH of 7.45 with 5 mM *β*-mercaptoethanol present to provide reducing conditions and therefore prevent protein oxidation. Attempts to concentrate beyond a maximum of 3 mg/ml resulted in the precipitation of SV3CP from solution and presumed binding of the precipitated protein to the membrane of the Centricon vessel. Despite the relatively low protein concentration of 3 mg/ml, crystallisation trails were attempted using all conditions of crystallisation screens from Jena Bioscience Single Hit Classics and Molecular Dimensions Crystal Screens I and II. In all instances the vapour diffusion hanging drop technique of protein crystallisation was used where the drop consisted of 2 μl of SV3CP solution mixed with an equal volume of crystallisation buffer. In the case of the Jena Bioscience screens the crystallisation drops were suspended over 0.7 ml of the crystallisation buffer, and with Molecular Dimension screens over 1 ml, dictated solely by the respective designs of the crystallisation screens. Following approximately two weeks, with crystallisation trays left at room temperature, a large number of crystals of varying quality and size were found in a number of conditions of the Jena Bioscience screen (the Molecular Dimensions screens yielded no protein crystals). Irrespective of crystallisation condition, all crystals, of all sizes, adopted a single morphology; that of a hexagonal column. To rationalise the number of crystals on which X-ray diffraction data would be collected, forty crystals were selected based upon appearance; well defined shape with sharp edges, and of suitable size (at the time a minimum of 20 μm in all dimensions) for data collection at a synchrotron radiation facility. To allow for data collection at cryogenic temperatures the crystals were allowed to equilibrate for 5 minutes in their respective crystallisation buffer containing the cryoprotectant glycerol at a concentration of 30% (glycerol was actually added in three equal volumes to the final concentration of 30% with a momentary pause between each addition to prevent possible crystal degradation) before being "fished" from the buffer/cryoprotectant solution using a mohair loop and flash frozen using a two step freezing process where the crystal is initially frozen in liquid ethane before being transferred for storage in liquid nitrogen.

4.4.2 Optimisation of native SV3CP crystallisation

Test diffraction images were taken using the rotating anode source at the University of Southampton, and based upon ability to diffract i.e. visual quality and resolution limit of a single test diffraction image, these incipient diffraction experiments identified a single condition; 10% w/v PEG 8000, 0.1 M HEPES pH 7.5 with 10% v/v ethylene glycol (Jena Bioscience Screen 4; condition: D5), that produced crystals of a superior quality (and incidentally largest in size) to any other crystal producing condition. Crystallisation optimisation experiments were subsequently completed over a further 330 conditions focussed around this initial hit. All buffer constituents were varied; PEG 8000: 5-15% w/v; 0.1M HEPES pH: 6.5 – 8.5; and ethylene glycol; 4 -12% v/v to produce the 330 permutations used in a grid screening approach. These optimisation experiments revealed a single condition; PEG 8000 7%, 0.1M HEPES pH 7.5 with 8% v/v ethylene glycol, that produced the highest quality, and largest, crystals on visual inspection (*Figure 4.18*).

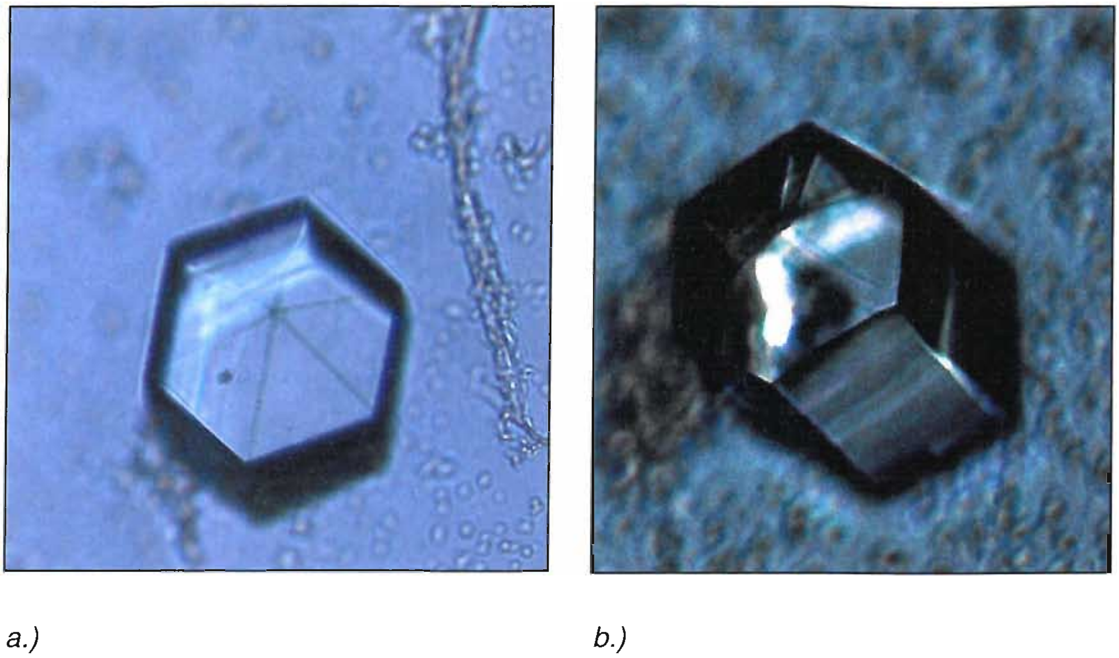


Figure 4.18: Crystals of native SV3CP grown using the vapour diffusion method at room temperature. *a.) Crystallisation condition: 10% w/v PEG 8000, 0.1 M HEPES pH 7.5 with 10% v/v ethylene glycol; crystal size: 80 x 20 x 20 µm. b.) Optimised crystallisation condition: 7% w/v PEG 8000, 0.1 M HEPES pH 7.5 with 8% v/v ethylene glycol; crystal size: 120 x 40 x 40 µm. Protein final concentration 1.5 mg/ml in drop in both cases.*

4.4.3 X-Ray diffraction data collection on native SV3CP crystals

A full dataset was subsequently collected on station ID14-1 (fixed wavelength beamline λ = 0.934 Å), ESRF, Grenoble, France, on a single crystal from the optimised condition over 180° of rotation, with an oscillation angle of 1° and a 1 second exposure time. Data reduction using MOSFLM for autoindexing and image integration followed by SCALA to scale and merge reflections revealed the dataset to be only of modest quality; d_{min}=2.9 Å and an R_{merge} of 14.5 (full data reduction statistics are listed in *Table 4.2*) .

Unit cell dimensions	a=129.5 Å b= 129.5 Å c= 119.7 Å
	α=90 β=90 γ=120

Space group	P6
Total number of reflections	269153
Number of unique reflections	14502
Resolution (Å)	2.9
Completeness (%)	100
R_{merge} (%) (outer resolution shell)	14.5 (34.9)
Multiplicity	11.5
Average I/σI (outer resolution shell)	3.3 (2.1)

Table 4.2: SCALA data reduction statistics for native SV3CP. *Outer shell values are given in parentheses.*

4.4.4 Phasing by molecular replacement

Unfortunately, further progress towards a structural solution with this initial dataset was prevented due to the lack of a suitable molecular replacement model structure from which initial phase estimates could be derived. No convincing peaks in rotation or translation function searches were seen when using the automated molecular replacement program MOLREP (163) with model structures of proteases related to the SV3CP by function; Poliovirus 3C protease (1L1N) (164), Type 2 Rhinovirus 3C protease (1CQQ) (165), Equine Arteritis Virus 3C protease (NSP4) (166), Hepatitis A Virus 3C protease (1HAV) (167) and Hepatitis A 3C protease in complex with an inhibitor (1QA7) (168). The putative molecular replacement model structures (those available July 2003) were identified by a simple SWISSPROT sequence identity search. With the highest degree of sequence identity at just 20% (Type 2 Rhinovirus 3C protease) it was of small surprise that a molecular replacement approach to determine initial phases would be unsuccessful.

4.4.5 Experimental phasing by multiple isomorphous replacement

In a logical progression in phase determination, phases had now to be sought by experimental means. The techniques of multiple isomorphous replacement (MIR) and single (SAD) and multi wavelength anomalous dispersion (MAD) have been widely applied and successful in phase determination. Firstly heavy metal soaks for MIR phase determination were attempted with ethylmercury chloride, a commonly used and successful mercury containing compound to generate heavy metal derivative protein crystals for use in MIR experiments. Ethylmercury chloride was added to final concentrations of 0.1 mM, 0.5 mM, 1.0 mM, 2.0 mM, 5.0 mM and 10.0 mM in a 4 µl drop of well solution (removed from a vapour diffusion experiment of conditions identical to the initial hit) containing the crystals of SV3CP from the optimised vapour diffusion experiments. Disappointingly all such attempts resulted in quite rapid degradation of the protein crystal.

4.4.6 Crystallisation of a selenomethionine derivative for phasing by Single / Multi-wavelength Anomalous Dispersion

Simultaneously to the MIR work, preparation of a selenomethionine derivative of the SV3CP began to allow for phasing by either SAD or MAD diffraction experiments. Full crystal trials were completed with the selenomethionine derivative SV3CP, however in line with expectations the physical properties of the derivative had been altered little from the native and the best quality crystals were obtained from identical crystallisation conditions as seen with the initial hit with native protein and possessed an identical morphology.

4.4.7 Collection of a MAD dataset and initial phasing attempts

A full MAD dataset was collected on a single crystal at the ESRF, Grenoble on station ID29 at cryogenic temperatures over three wavelengths identified by fluorescence scan prior to data collection; peak (λ = 0.97932 Å; 270° of data; 1° oscillation angle, 1 second exposure), inflection point (λ = 0.97915 Å; 180° of data; 1° oscillation angle, 1 second exposure) and high energy remote (λ = 0.97702 Å; 180° of data; 1° oscillation angle, 1 second exposure). Initial data processing to SCALA showed data to be of relatively low quality; peak data set: d_{min} = 3.3 Å, R_{sym} = 17.7 %; inflection point data set: d_{min} = 3.9 Å, R_{sym} = 16.7%; remote data set: d_{min}= 3.9 Å, R_{sym} = 15.7%. Unit cell parameters and crystallographic symmetry were identical to crystals of native protein. Had the MAD data constituted the entire diffraction data collected to obtain a structural solution it would have been of too poor quality to warrant further attention. However, as a means to determine phase information that could be applied to the native data set, previously collected (d_{min} = 2.9 Å) it was perfectly adequate. Hence, attempts were subsequently made to spatially locate anomalous scattering atoms and therefore derive phase information by use of the phasing package of SOLVE and RESOLVE. For a realistic chance of experimentally phasing by MAD techniques, a minimum of one selenomethionine residue per one hundred residues is recommended with a maximum of fifty anomalously scattering atoms in the asymmetric unit (although these guidelines are subject to change as phasing programs improve). Each molecule of SV3CP contains five methionine residues, therefore in the selenomethionine derivative; five selenomethionines. The selenium atom contained in each selenomethionine residue is a point of anomalous scattering. With five methionines in 181 residues phasing requirements were comfortably met. Gel filtration analysis had previously provided strong evidence that SV3CP existed as a dimer when a biologically functioning unit; indeed, cell content analysis by application of the Matthews co-efficient indicated a cell solvent content of 62.13% and two molecules per asymmetric unit. Therefore the location of at least five but almost certainly ten selenium sites were being sought by use of the SOLVE/RESOLVE package. Unfortunately all attempts were unsuccessful. SOLVE utilises a system of four scoring criteria; analysis of difference Pattersons, self difference Fourier analysis, a non randomness test on a native Fourier and figure of merit of phasing (a fuller explanation is offered in *Section 2.4.4.6*) to arrive at a single statistical measure, the Z score. The Z score reflects the reliability of a solution defining the locations of a set of anomalously scattering atoms from which phases can be derived. A promising solution should have a Z score exceeding a threshold minimum of 20, though a higher Z score would be expected with larger numbers of anomalous scattering atoms without being reflective of an improvement in solution. In addition the phasing figure of merit should have been above 0.6. Repeated attempts to locate anomalously scattering atoms with this MAD data resulted in Z scores not exceeding 12, and phasing figures of merit no greater than 0.4. Further to the statistical verification of the absence of a suitable solution, comparison of positions of anomalously scattering atoms between separate runs of SOLVE revealed little continuity, with different sets of sites being picked on each run. However, it should be noted that although the variation in anomalous scattering atom sites seen initially may have indicated an unreliable solution; between separate runs of SOLVE, sites may be found not simply within a single protein molecule but also in symmetry-related molecules. Further suspicion of the reliability of the solution was raised when sites were frequently seen lying along cell axis. Such sites are notoriously spurious as noise is often seen in Patterson maps (used by SOLVE to locate anomalously scattering atoms) along cell axis. Finally, and most condemningly, inspection of electron density maps calculated with phases determined by SOLVE based upon these suspect sites, showed little, if anything, in decipherable protein secondary structure with contiguous electron density and no indication of a protein / solvent boundary.

4.4.8 A flexible protein?

At this stage it was proposed that in its present state, without any peptide bound, the SV3CP may be exhibiting some flexibility hence the relatively low quality of diffraction data obtained thus far and the inability to decipher definite sites of anomalously scattering atoms. However, in complex with a peptide or other compound capable of locating in the SV3CP active site cleft, the putatively flexible protein may be held in a more rigid state. Accordingly efforts were made towards the design and synthesis of a compound capable of strongly binding within the active cleft and therefore bracing the SV3CP. The four chromogenic peptides that had already been synthesised proved of considerable use in judging the activity of recombinantly expressed SV3CP. The peptides consisted of a peptide portion of between three and six residues in length mimicking the sequence leading up to the most preferentially cleaved site of the five sites of cleavage within the natural peptide, namely the polyprotein product of ORF1 of the SV genome. At the *C*-terminus of the peptide was placed a chromogenic group that upon cleavage from the peptide portion yielded a strong yellow colour measurable spectrophotometrically at 405 nm. Since any substrate, by definition, would be turned over by SV3CP and therefore only transiently occupy the active site the likelihood of crystallising a complex of SV3CP with substrate would be considerably slimmer than with a compound that irreversibly bound to the active site cleft. The initial project aims had been to solve the structure of SV3CP, followed by the design of a knowledge based compound capable of inhibiting SV3CP activity making it suitable for development as a therapeutically viable agent capable of tackling infection by SV. However, since the crystallographic work was proving unsuccessful the design of the inhibitor would have to be based solely upon the reaction kinetics studies completed with the chromogenic peptides and not be knowledge based as had been hoped for. It was clear from the kinetics studies that, of the four chromogenic peptides the peptide containing the peptide sequence: EFQLQ, was cleaved at a greater rate than the other three: QLQ, FQLQ, DEFQLQ. It was a logical progression therefore to take the EFQLQ peptide as the basis and therefore means of locating a peptidyl inhibitor within the active site of the SV3CP. Further, by replacing the *C*-terminal chromophore of the peptide by a reactive group capable of covalently bonding to the active site cysteine of the SV3CP would, if successful, provide not only a means of bracing the SV3CP for the benefits of the crystallographic work but also achieve the aim of developing a compound capable of arresting SV3CP activity and ultimately SV replication.

When in close proximity to the side chain sulfhydryl group of cysteine residues, Michael acceptor type groups may bond covalently to the sulfhydyl group sulphur by acting as a soft electrophile. In pursuit of an effective inhibitor of the SV3CP, an inhibitory compound was designed that adopted the peptide sequence: EFQLQ as with the chromogenic peptide, but in place of the chromophore was attached a Michael acceptor group in the form of a *C*-terminal propenyl-ethyl-ester extension. The overall design rationale was that the penta-peptide portion would serve to locate the inhibitor tightly within the active site cleft leaving the functional portion: the Michael acceptor *C*-terminal extension, in close proximity to the active site cysteine side chain. This would allow the inhibitor to covalently bind to the sulfhydryl group sulphur, rendering the SV3CP irreversibly inhibited. Fortunately, not only was the rather involved synthesis of the inhibitor a success but also was its efficacy as a potent inhibitor of the SV3CP. For a more detailed explanation of the inhibitor work see *Sections 3.5 - 3.11*.

4.4.9 Modification of SV3CP with MAPI

4.4.9.1 Achieving SV3CP at high concentrations

Up to this stage SV3CP had been concentrated to a maximum concentration of 3 mg/ml, a relatively low protein concentration for protein crystallography. Concentration above 3 mg/ml had resulted in the aggregation and precipitation of SV3CP. To maximise the chances of successful crystallisation of a SV3CP-MAPI complex it was apparent it would be necessary to achieve higher concentration SV3CP solutions. One method of increasing protein solubility is the phenomenon of 'salting in' where increasing the strength of the ionic environment by the addition of a salt keeps protein in solution through binding of divalent salt ions to the protein surface resulting in a reduction in the preferred level of hydration of the protein. To these ends a series of experiments were completed where NaCl was added to 1 ml of SV3CP at 2 mg/ml in its phosphate purification buffer to final concentrations of: 50 mM, 100 mM, 150 mM, 200 mM, 250 mM, 300 mM, 350 mM, 400 mM, 450 mM, and 500 mM. Each of these protein samples were then concentrated using Centricon YM-10 concentration vessels. As a means of tracking the solubility limit of SV3CP in each of the NaCl concentrations, protein concentration was measured at intervals during concentration by use of BIORAD protein concentration assay. The intervals at which protein concentration was measured were after the sample size of each sample had reduced from its starting volume of 1 ml to a volume of: 500 µl, 250 µl, 100 µl, and finally 50 µl.

	Protein concentration (mg/ml) at sample volume of:			
[NaCl]/ mM	500μl	250μl	100μl	50μl
50	2.8	n/a	n/a	n/a
100	2.6	n/a	n/a	n/a
150	2.8	3.2	n/a	n/a
200	2.8	3.2	n/a	n/a
250	3.1	5.3	9.0	11.2
300	3.8	7.4	13.7	18.2
350	1.1	n/a	n/a	n/a
400	n/a	n/a	n/a	n/a
450	n/a	n/a	n/a	n/a
500	n/a	n/a	n/a	n/a

Table 4.3: Solubility of SV3CP dependent upon NaCl concentration*. Where result is indicated by 'n/a' BIORAD assay showed no increase in absorbance at 595 nm above that of the blank sample; protein may have been present in solution though at concentrations too low to measured.*

The results (*Table 4.3*) clearly showed the need for a relatively high concentration of NaCl; 300 mM to provide a sufficiently strong ionic environment to keep SV3CP in solution at concentrations above 3 mg/ml. Further, NaCl at concentrations of 350 mM or above resulted in the opposite phenomenon of 'salting in' by causing 'salting out' i.e. to cause precipitation of a soluble protein by increasing the strength of the ionic environment beyond that leading to increased solubility. Of course these observations may only hold true for the SV3CP in a phosphate buffer of pH 7.45 at room temperature (pH and temperature being two factors that can have a large effect upon protein solubility). In summary, it was now possible to concentrate the SV3CP up to a concentration of 20 mg/ml if in the presence of NaCl at 300 mM, a significant if simple step to improving chances of crystallisation.

4.4.9.2 Modification of SV3CP with MAPI whilst maintaining solubility

The crystallographic purpose of the MAPI was to brace the SV3CP to increase its rigidity and assist in the formation of more regularly packed crystals to improve diffraction data quality. Therefore the approach of soaking the MAPI into crystals previously obtained of SV3CP that had already proven to be of inadequate quality, was not appropriate. The SV3CP and MAPI would therefore have to be in complex before use in crystal trials.

Initial attempts to form a complex of SV3CP with MAPI by simply dissolving a three fold molar excess of MAPI in an aqueous solution of SV3CP proved unsuccessful due to the insoluble nature of the inhibitor. MAPI had to this point proved so insoluble that during its synthesis it was necessary to dissolve it in 100% DMSO prior to loading onto the C18 reverse phase column for purification. The converse was true of SV3CP; typical of non membrane bound proteins SV3CP is relatively hydrophilic. Therefore it was necessary to provide a single buffer that provided the necessary environment for dissolution of both SV3CP and MAPI. It was important to ascertain how hydrophobic in nature a solution may be before causing SV3CP to aggregate and precipitate. Addition of the organic solvent DMSO to a protein solution would cause an increase in its hydrophobic nature. To these ends DMSO was added to a series of SV3CP samples (15 mg/ml concentration) to the final concentrations of; 1%, 2%, 5%, 10%, 20%, 30%, 40% and 50% (solubility assays were completed in duplicate). Protein aggregation was then judged spectrophotopically by measuring optical density at 600nm. With an entirely soluble and non-aggregated protein sample no scattering of light of 600 nm in wavelength should be seen. In contrast significant scattering would be seen at this wavelength in a protein sample containing aggregated protein. In cases of severe protein aggregation the solution would become visibly cloudy.

DMSO	1%	2%	5%	10%	20%	30%	40%	50%
Precipitation by eye	No	No	No	No	No	Moderate	Quite heavy	Heavy
Absorbance at 600nm (duplicate readings shown)	0.065 0.040	0.024 0.000	-0.014 -0.014	-0.016 -0.016	-0.018 -0.016	0.163 0.167	0.373 0.412	0.221 0.213
Average absorbance at 600nm	0.0525	0.012	0.000	0.000	0.000	0.165	0.3925	0.217

Table 4.4: Solubility of SV3CP as a function of increasing DMSO concentration.

At concentrations of 30% DMSO and greater, protein aggregates formed therefore such conditions were not suitable. At 20% DMSO there was no evidence of aggregation, suggesting aggregation would occur only in solutions consisting of more than 20% DMSO. To remain well within this limit a three molar excess of MAPI over SV3CP was dissolved in a volume of DMSO so that on addition to SV3CP in phosphate buffer would constitute only 10% of the total final solution volume. Before addition to the SV3CP the solution of MAPI in DMSO was divided into 10 aliquots of equal volume then added as follows; 5 aliquots together followed by 10 minutes incubation at room temperature, followed by 5 additions of a single aliquot each separated by same incubation period of 10 minutes. The success of this approach was not altogether unexpected; a similar method of dissolving the MAPI in 100% DMSO before transferring into an aqueous buffer had been used with success during the synthesis and purification of MAPI. To achieve a homogenous sample, usually essential for protein crystallisation, excess MAPI was separated from the SV3CP-MAPI complex by passing the reaction sample through a micro-centrifuge column (usually used in DNA preparation work) filled with G25 Sephadex media and using the standard phosphate buffer (+300 mM NaCl) to elute the pure SV3CP-MAPI complex. This rudimentary purification step has the additional benefit of removing DMSO from the SV3CP-MAPI sample. Finally, it should be noted that for the successful preparation of a sample of SV3CP in complex with MAPI at concentrations of up to 20 mg/ml it is essential to concentrate the SV3CP first, then complex with MAPI and finally gel filtrate. If MAPI is added to a solution of SV3CP of low concentration and then concentrated the complex is lost from solution, judged by BIORAD protein assay, presumably due to binding to the concentration vessel membrane. The reasons for this remain unclear.

4.4.9.3 Confirmation of SV3CP modification by MAPI

Initial assessment of SV3CP inhibition was made by incubation of 5 µl of the putative SV3CP/MAPI complex at 15 mg/ml, with 5 µl of a 100 mg/ml solution of the chromogenic peptide Ac-EFQLQ-*p*NA. Following 15 minutes incubation at room temperature the absence of the visible development of any yellow colour (corresponding to cleavage of the peptide) was indicative of a completely inactivated SV3CP sample. Due to the extremely small assay volume it was not possible to ascertain quantitatively any peptide cleavage in a standard laboratory spectrophotometer; therefore this approach was only ever used qualitatively. It should be re-enforced that prior to the use of any SV3CP, activity was always confirmed by assay with the chromogenic peptide Ac-EFQLQ-pNA with the development of a strong yellow colour seen within 2 -3 minutes.

Further confirmation and accurate assessment of MAPI binding was completed by ESI-oa-TOF-MS. A single major peak seen on the spectrum (*Figure 4.19*) at 20045 Da corresponded to one molecule of SV3CP plus one molecule of MAPI indicating that all molecules of SV3CP had been irreversibly modified by binding MAPI. A single peak was also indicative of each molecule of SV3CP binding only a single molecule of MAPI. A single molecule of SV3CP possesses 5 cysteine residues, each is theoretically capable of reacting with MAPI. Though this would be unlikely since the MAPI had been specifically designed to include a peptide substrate sequence to bring the Michael acceptor portion into close proximity to the active site cysteine; without the peptide portion it would be unlikely the functional Michael acceptor *C*-terminal extension would be bought into close enough proximity for the quite subtle nucleophilic attack on the MAPI propenyl-ethyl-ester moiety by the cysteine side chain sulfhydryl.

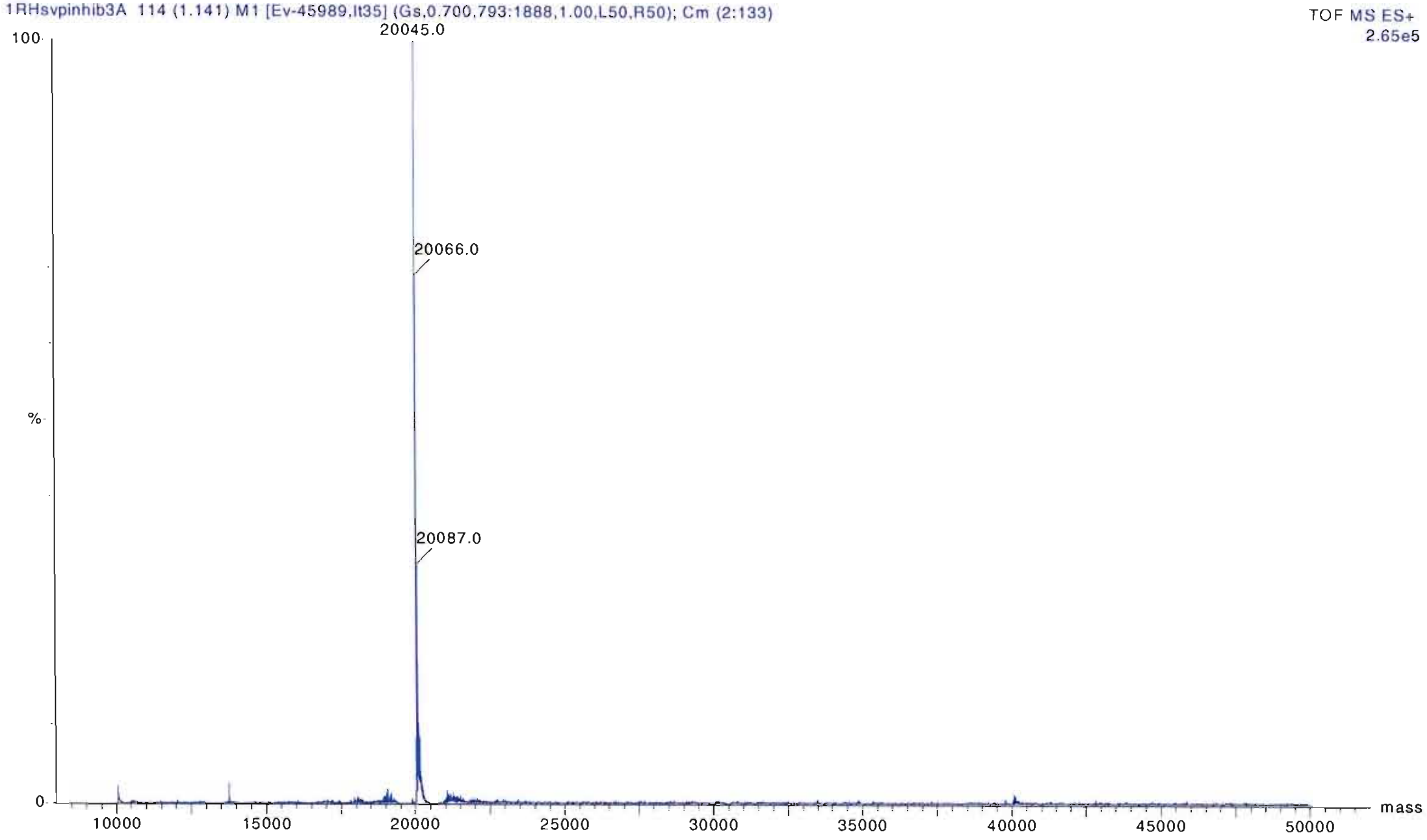


Figure 4.19: ESI-oa-TOF mass spectra of native SV3CP with covalently bound MAPI. *A single main peak corresponding to the exact calculated mass (20045 Da) of a single molecule of SV3CP with a single molecule of MAPI bound was observed.*

4.4.9.4 Crystallisation trials of SV3CP in complex with MAPI

SV3CP modified by covalently bound MAPI, represented a species of sufficiently different character from the original native SV3CP therefore crystallisation screens were repeated employing the hanging drop vapour diffusion technique (drop size: 2 µl SV3CP-MAPI sample + 2 µl crystallisation buffer; protein stock solution of 15 mg/ml therefore 7.5 mg/ml in the drop) using the full complement of crystallisation conditions provided by the Jena Bioscience Single Hit Classics kit. Further, a novel crystallisation condition was actively sought if crystals of improved quality were to be grown. Following two weeks of incubation at room temperature, the crystal trays were checked for evidence of crystal growth. Crystals of varying morphologies, size and quality were observed across a number of conditions. Crystallisation trial results are summarised in *Table 4.5*.

JB Screen	Well no.	Temp / °C	Condition	[Protein] mg/ml	Crystal description
4	A1	24	25% PEG 5000 MME, 100 mM TRIS-HCl pH 8.5, 200 mM Li-sulphate	15	3 large (approx 300 x 300 µm), thick plate-like.
3	B2		15% w/v PEG 4000, 100 mM Na-Citrate pH 5.6, 200 mM Ammonium Sulphate	15	20 large crystals, approx 1.5 mm x 200 µm wide; depth approximately 100 µm.
3	C1		20% PEG -4000, 20% w/v Iso-propanol, 100 mM Na-Citrate	15	Single large 250 x 350 x 200 µm; layered crystal but with relatively sharp edges.
10	B6		1.5 M Li-Sulphate, 100 mM TRIS-HCl pH 8.5	15	Microcrystal shower –very small but clean looking crystals.
1	C2		25% w/v PEG 550 MME, 100 mM Na-MES pH 6.5, 10mM Zn-Sulphate	15	Single medium sized crystal 80 x 80 x 100 µm- diamond shaped – average quality.
8	D1		30% Ethanol w/v , 10 % PEG 6000 w/v, 0.1 M Na-acetate	15	Single large 500 x 400 x 200 µm crystal, but very layered.
7	C1		30% MPD, 100 mM Na-MES pH 6.5, 200mM Mg-Acetate	15	Approx 15 small crystals 50 x 50 x 20 µm.
3	C6		25% w/v PEG 4000, 100mM Na-Citrate pH 5.6, 200 mM Ammonium Sulphate		Approx 20 tiny needle like crystals 300 x 10 x 10 µm; unusable for diffraction experiment.

Table 4.5: Excerpt of crystallisation screen results of SV3CP modified with the MAPI.

Crystals were frozen from each of the conditions described in Table 4.5 producing useable crystals (in an identical manner as with crystals of the free form of SV3CP) and single test diffraction images were taken for each crystal on station ID-23-1 at the ESRF, Grenoble. Visual inspection of each test image revealed crystals from the Jena Bioscience condition JB4 A1; 25% PEG 5000 MME, 100 mM TRIS-HCl pH 8.5, 200 mM Li-sulphate, produced the highest resolution data with diffraction to approximately 1.8 Å that was also of the highest quality with tight, non-overlapped spherical reflections (*Figure 4.20*).

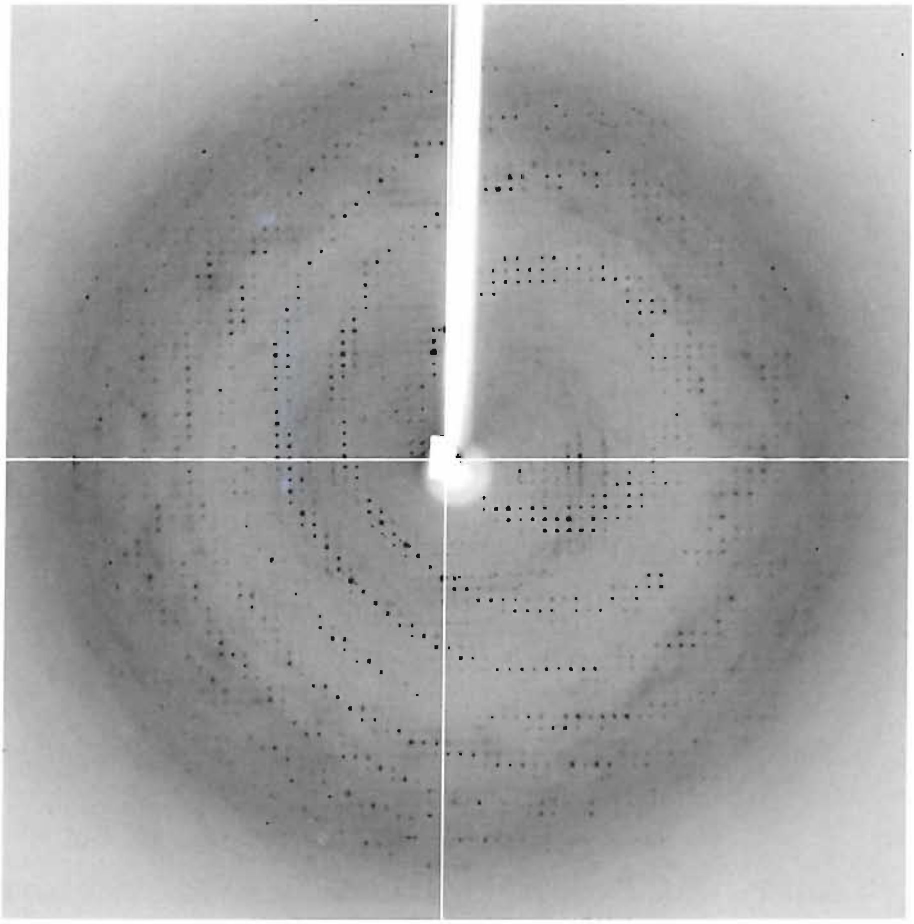


Figure 4.20: Typical diffraction pattern obtained during data collection on a crystal of SV3CP modified with MAPI. *Crystals were grown in the JB1 A1 condition and data was collected on beam-line ID14-1, ESRF over 180 ° with 1 ° oscillation angle and 1 s exposure.*

4.4.9.5 X-Ray diffraction data collection with native SV3CP-MAPI complex crystals

A full dataset (180° of data; 1° oscillation angle, 1 second exposure) was collected from a single crystal from condition JB4 A1 on station ID14-1 at the ESRF. Cell parameters and space group were determined and images integrated using MOSFLM then scaled and merged using SCALA. SCALA statistics are summarised in *Table 4.6*.

Cell parameters	a = 49.49 b = 84.10 c = 121.47 $\alpha = \beta = \gamma = 90.000$
Space group	P 2_1 2_1 2_1
Wavelength (Å)	0.93400
Overall resolution range (Å)	40.49 -1.79
High resolution range (Å)	1.75 - 1.70
R_{merge} (%)	0.053 (0.623)
Total number of observations	376358 (38859)
Total number of unique reflections	55934 (7504)
Mean (I/ σ (I))	22.4 (2.4)
Completeness (%)	98.8 (92.8)
Multiplicity	6.7 (5.2)

Table 4.6: SCALA data reduction statistics for native SV3CP modified by MAPI. *Outer shell values are given in parenthesis.*

With a resolution limit of 1.75 Å and overall R_{merge} of 5.3% the quality of data collected on this crystal form of the SV3CP-MAPI complex represented diffraction data of high resolution and quality, and a large improvement on diffraction data obtained from the free form of the SV3CP. The MAPI had fulfilled its role in aiding conformational stability of SV3CP for crystallisation.

Some ambiguity existed over the exact space group of the crystal on which the data was collected prior to image integration with MOSFLM. Following recommended protocol data is best processed through MOSFLM in the highest symmetry with the lowest penalty but without any additional symmetry operators e.g. it is best to process data in P2 and not $P2_1$. Any ambiguity over space group was easily resolved by inspection of the truncated reflection file following scaling and merging. Pseudo-precession images (visual representations of zones within reciprocal space), derived from the truncated reflection file, were displayed in HKLVIEW (115). Inspection of images laying on the *hk*0, *h*0*l* and 0*kl* planes of *reciprocal* space revealed tell tale systematic absences of space group $P2_12_12_1$ running along all three axes; *h*, *k* and *l*. The space group $P2_12_12_1$ possesses what are termed 'screw axes' along all three axes of *real* space. A screw axis is a symmetry operator that describes the rotation about a cell axis in real space and translation along that axis of protein molecules in relation to each other within a crystal system. In the case of a $P2_1$ space group, a protein molecule will be found 180° rotated from a protein molecule laying at the axial origin, and translated by ½ a cell length away along. In a $P2_12_12_1$ space group this symmetry operator is replicated in all three cell axes: a, b, and c. A visual manifestation of this can be seen by observing systematic absences along the *h*, *k* and *l* axes in *reciprocal* space. For a $P2_12_12_1$ space group every even numbered reflection is present and every odd numbered reflection is absent along each of the three reciprocal space axes.

4.4.9.6 Crystals of SV3CP-MAPI complex suffer degradation

It should be noted that crucial to achieving high quality data with this crystal form was the freezing of crystals within a week of their appearance; this was usually between 18 and 21 days following the setting up of the crystallisation experiments. If this time limit was exceeded crystals were clearly seen to degrade, presumably as conditions within the sealed environment of a crystal tray, proceed beyond the appropriate conditions for crystal growth. This phenomenon was observed repeatedly (*Figure 4.21*).

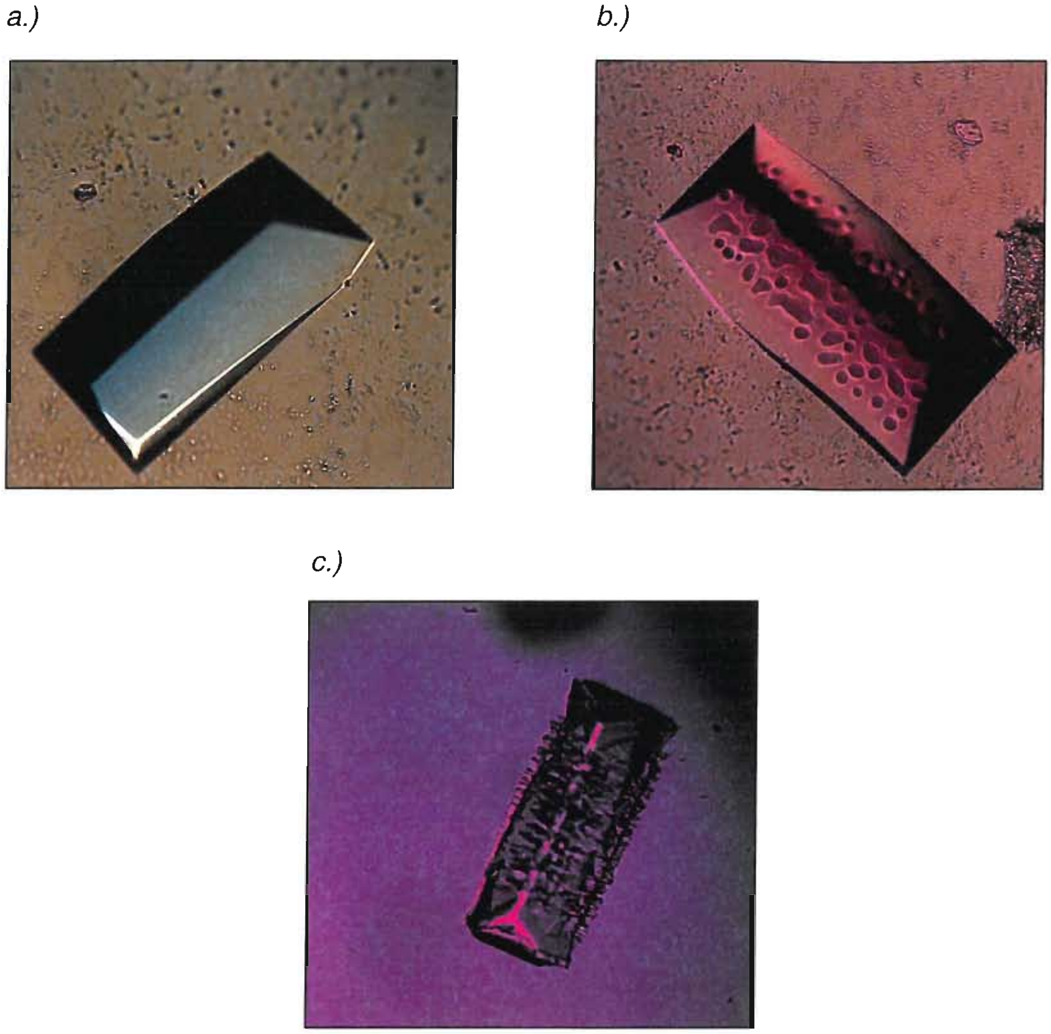


***Figure 4.21: Degradation of crystals of SV3CP modified with MAPI**. a) Crystal in hanging drop at 18 days; b) at 23 days; c) at 28 days.*

4.4.9.7 Optimisation of SV3CP-MAPI complex crystallisation

Optimisation experiments to determine improved crystallisation conditions for the SV3CP-MAPI complex were completed adopting a grid screening approach with all constituents of the crystallisation buffer of the original hit varied as described in *Table 4.7*.

	pH					
% PEG 5000 MME	7.0	7.5	8.0	8.5	9.0	[LiSO$_4$] / mM
15%	**1.** 15% , 0	**2.** 15% , 0	**3.** 15% , 0	**4.** 15% , 0	**5.** 15% , 0	0
	6. 15%, 200	**7.** 15%, 200	**8.** 15%, 200	**9.** 15%, 200	**10.** 15%, 200	200
20%	**11.** 20%, 0	**12.** 20%, 0	**13.** 20%, 0	**14.** 20%, 0	**15.** 20%, 0	0
	16. 20%, 200	**17.** 20%, 200	**18.** 20%, 200	**19.** 20%, 200	**20.** 20%, 200	200
25%	**21.** 25%, 0	**22.** 25%, 0	**25.** 25%, 0	**26.** 25%, 0	**27.** 25%, 0	0
	23. 25%, 200	**24.** 25%, 200	**28.** 25%, 200	**29.** 25%, 200	**30.** 25%, 200	200
30%	**31.** 30%, 0	**32.** 30%, 0	**33.** 30%, 0	**34.** 30%, 0	**35.** 30%, 0	0
	36. 30%, 200	**37.** 30%, 200	**38.** 30%, 200	**39.** 30%, 200	**40.** 30%, 200	200
35%	**41.** 35%, 0	**42.** 35%, 0	**43.** 35%, 0	**44.** 35%, 0	**45.** 35%, 0	0
	46. 35%, 200	**47.** 35%, 200	**48.** 35%, 200	**49.** 35%, 200	**50.** 35%, 200	200

Table 4.7: Optimisation strategy for obtaining improved crystals of the SV3CP-MAPI complex. *Shown are details of the crystallisation conditions used in crystal optimisation experiments in search of improved X-ray diffraction data for the original condition producing high quality crystals. 25% PEG 5000 MME, 100mM TRIS-HCl pH 8.5, 200mM Li-sulphate.*

Whilst crystals formed in a number of conditions in the optimisation trial, none were quite as impressive in appearance or size as those seen from the original hit. Later test images to determine diffraction resolution and quality on station ID-23-1 at the ESRF, Grenoble, failed to identify a crystal that gave improved diffraction over crystals from the original hit. It was clear that in all further crystallisation of the SV3CP-MAPI complex the original crystallisation condition would be used exclusively.

With a reliable protocol to complex the SV3CP and MAPI established and crystallisation conditions identified that could routinely produce crystals giving rise to high quality diffraction data, it was now necessary to obtain SAD / MAD data from a crystal of selenomethionine derivative SV3CP in complex with MAPI.

4.4.9.8 Co-crystallisation of a selenomethionine derivative SV3CP and MAPI

As with crystallisation of the free form of SV3CP, the selenomethionine derivative SV3CP when in complex with MAPI crystallised in identical conditions as to native SV3CP in complex with MAPI, i.e. using protein at a stock concentration of 15 mg/ml (therefore 7.5 mg/ml in the drop) and a crystallisation buffer of 25% PEG 5000 MME, 100 mM TRIS-HCl pH 8.5, 200 mM Li-sulphate at room temperature. A combination of factors prevented full crystallisation screens or optimisation experiments being completed with the selenomethionine derivative SV3CP modified with MAPI. Firstly large scale crystallisation screens can be impractical with selenomethionine protein due to its relative expense to produce. Secondly a high quality dataset had already been obtained from native SV3CP in complex with MAPI therefore the information being sought from the selenomethionine derivative crystals was to allow for phasing only; not that this information is trivial to obtain. Maximal anomalous differences are measurable at resolutions of around 3 Å, therefore a *reasonable* quality crystal generated from the native crystallisation condition would be perfectly adequate.

4.4.9.9 MAD data collection on SV3CP-MAPI complex crystals

A full MAD dataset, over three wavelengths was collected on station ID-23-1 at the ESRF, Grenoble, France. In a rather impressive feat of endurance a single crystal stood up to a fluorescence scan prior to diffraction data collection followed by collection of 540° of diffraction data at the peak wavelength, with subsequent collection of 180° of diffraction data at the inflection point wavelength and a low energy remote wavelength. In total this was a quite remarkable 900° of data from a single crystal. At all wavelengths data was collected employing a 1° oscillation angle and 1 second exposure.

4.4.9.10 Fluorescence scan to determine data collection wavelengths

Theoretical values for anomalous scattering near the absorption edge of an anomalously scattering atom are not accurate due to scattering fluctuations as a result of the selenium environment i.e. within the protein crystal. Therefore energy values for the incident X-ray beam for peak, inflection point and remote dataset collection must be ascertained experimentally. Completion of a fluorescence scan on the candidate crystal allows X-ray fluorescence to be measured as a function of the incident X-ray energy.

The X-ray beam energy (and hence wavelength) where maximal anomalous scattering occurs can be determined from the fluorescence scan curve; this value is termed f" max, and the dataset collected at this energy is called the peak dataset. The peak dataset will contain maximal Bijovet differences i.e. measurable differences *within* a dataset between reflection intensity of Friedel pairs as a result of anomalous scattering. Bijovet differences can be exploited to estimate selenium atom positions and allow phase information to be derived, and ultimately extended to, protein atoms. Despite Bijovet differences being only relatively small, the order of a few percent (measurement comprises much of the difficulty of anomalous dispersion techniques), a peak dataset contains sufficient information to phase successfully by SAD; a phasing technique reliant solely upon Bijovet differences. However to maximise the chances of experimental phasing it is prudent to collect two further datasets; an inflection point dataset (also termed; edge dataset) and a remote dataset. The normal component of atom scattering, f' is reduced to its lowest possible value at the inflection point X-ray energy. A third dataset, called the remote dataset, is collected at an X-ray energy where f' is close to its normal value and f" is dependent upon whether a high or low energy remote is collected. Energy values for f" and f' at peak (f" maximum), inflection point (f' minimum) were calculated and associated wavelength identified using the program CHOOCH from the fluorescence scan spectra collected. Based upon beam and crystal properties CHOOCH suggested an accurate wavelength to which to tune the X-ray beam to achieve energies corresponding to peak and inflection point wavelengths; the remote wavelength was selected manually as the need for an accurate X-ray energy is less. In this instance a high energy remote was chosen. Details of energies, wavelengths and f' and f" values are shown in *Table 4.8.*

Dataset	Energy / keV	Wavelength / Å	f'	f"
Peak (f" max)	12.661	0.97925	-4.90000010	4.69999981
Inflection point (f' min)	12.657	0.97955	-8.89999962	2.29999995
High energy remote	12.699	0.97625	-4.00000000	4.00000000

Table 4.8: X-ray beam parameters used during the collection of MAD data over three wavelengths*. F' and F" values are shown for the peak (f" max), inflection point (f' min), and a high energy remote energies.*

4.4.9.11 Data reduction of SV3CP-MAPI MAD data

Auto-indexing to determine cell parameters and space group along with diffraction image integration was completed using MOSFLM (116). The data set was then scaled using SCALA (115) but reflections were not merged, this is essential in order to maintain the vital intensity variations between Friedel pairs. SCALA statistics covering the three datasets are listed in Tables: *4.9*, *4.10* and *4.11.* Note that crystal cell parameters differ very slightly between datasets. Since all data was collected from a single crystal this would not be expected but can be accounted for through inherent data collection inaccuracies; crystal slippage, beam intensity fluctuation and is not of any real concern. Besides, following SCALA the three datasets were merged into a single reflection file using SCALEIT (115) with inflection point and remote data being normalised to the peak dataset cell parameters.

Cell parameters	a = 49.69 b = 85.15 c = 120.88
	α = β = γ = 90.000

Space group	P $2_1 2_1 2_1$
Wavelength (Å)	0.97925
Overall resolution range (Å)	40.46 - 2.26
High resolution range (Å)	2.25 - 2.20
R_{merge} (%)	0.086 (0.741)
Total number of observations	521622 (34916)
Total number unique	26649 (1938)
Mean (I/σ(I))	29.3 (3.5)
Completeness (%)	99.5 (98.7)
Multiplicity	19.6 (18.0)
Anomalous completeness (%)	99.4 (97.9)
Anomalous multiplicity	10.3 (9.4)

Table 4.9: Peak wavelength dataset. *SCALA data reduction statistics for selenomethionine derivative SV3CP modified with MAPI. Outer shell values are given in parentheses.*

Cell parameters	a = 49.79 b = 84.91 c = 120.93 $\alpha = \beta = \gamma = 90.000$
Space group	$P\ 2_1\ 2_1\ 2_1$
Wavelength (Å)	0.97955
Overall resolution range (Å)	46.03 - 2.50
High resolution range (Å)	2.56 - 2.50
R_{merge} (%)	0.057 (0.137)
Total number of observations	120862 (8552)
Total number of unique reflections	18029 (1332)
Mean ($I/\sigma(I)$)	22.8 (10.5)
Completeness (%)	98.2 (99.9)
Multiplicity	6.7 (6.4)
Anomalous completeness (%)	97.7 (98.6)
Anomalous multiplicity	3.5 (3.3)

Table 4.10: Inflection point dataset. *SCALA data reduction statistics for selenomethionine derivative SV3CP modified with MAPI. Outer shell values are given in parentheses.*

Cell parameters	a = 49.74 b = 84.62 c = 120.79
	α = β = γ = 90.000

Space group	P 2_1 2_1 2_1
Wavelength (Å)	0.97625
Low resolution range (Å)	49.15 - 2.50
High resolution range (Å)	2.56 - 2.50
R_{merge} (%)	0.064 (0.158)
Total number of observations	117483 (8314)
Total number of unique reflections	17803 (1320)
Mean (I/σ(I)	20.7 (8.4)
Completeness (%)	97.6 (100.0)
Multiplicity	6.6 (6.3)
Anomalous completeness (%)	97.0 (98.7)
Anomalous multiplicity	3.5 (3.2)

Table 4.11: Remote wavelength dataset*. SCALA data reduction statistics for selenomethionine derivative SV3CP modified with MAPI. Outer shell values are given in parentheses.*

All three datasets showed an adequate resolution of data to allow for MAD phasing; peak d_{min} = 2.20 Å, inflection point d_{min} = 2.5 Å, and remote d_{min} = 2.5 Å. Overall data quality for all three datasets was high reflected in R_{merge} values of 8.6%, 5.7% and 6.4% for the peak, inflection point and remote datasets respectively. A further measure of overall data quality is overall data completeness and neared 100% for all three datasets. Early indications of suitability of the data for MAD phasing are the SCALA dataset statistics; overall multiplicity and anomalous multiplicity. To increase the chances of successfully determining the subtle reflection intensity differences between Friedel pairs it is desirable to have a dataset of high redundancy; i.e. the same reflections recorded a number of times; hence 540 ° of data were collected for the peak dataset when 180 ° would have been sufficient for a native dataset. The overall multiplicity of a dataset is indicative of dataset redundancy; at 19.7, 6.7 and 6.6 for peak, inflection point and remote datasets multiplicity could be considered high. The anomalous multiplicity was high for the peak dataset at 10.3, reflective of the large amount of data collected and fair for the remaining datasets; the inflection point and remote both at 3.5.

Following SCALA, reflection intensities were converted into structure factor amplitudes using TRUNCATE (115).

4.4.10 Using SOLVE to phase by MAD

The three datasets were prepared for use with SOLVE using SCALEIT to scale and merge the inflection point and remote datasets to the peak dataset. Inspection of the SCALEIT log file revealed little of interest with the exception that the new reflection file contained 26622 unique observations in the resolution range 46.029 – 2.188 Å. The merged reflection file was then passed to SOLVE for anomalous atom site location and subsequent initial phase estimations for the anomalous scattering atoms to be made that are then extended to protein atoms. SOLVE is quite autonomous requiring little user input other than definitions of the f' and f", and wavelength for each of the original datasets; peak, inflection point and remote, plus the resolution range over which to include data for phasing, in this case 27.1 - 3.0 Å using 26622 reflections.

SOLVE took a total of 17 minutes to find a promising solution, managing to locate all 10 selenium atoms corresponding to the 10 selenomethionines, 5 per molecule, of the putative SV3CP-MAPI dimer (*Table 4.12*).

```
------TOP SOLUTION FOUND BY SOLVE  (<m> = 0.56; score =  26.33) --------

        X      Y      Z        OCCUP     B          HEIGHT/SIGMA

  1   0.392  0.806  0.134    1.060   43.6          14.8
  1   0.435  0.731  0.016    1.131   46.5          11.9
  1   0.893  0.717  0.240    1.216   60.0          10.5
  1   0.363  0.934  0.146    0.865   60.0           9.6
  1   0.678  0.779  0.232    0.492   37.0           8.3
  1   0.402  0.153  0.232    0.789   48.9           6.5
  1   0.573  0.606  0.095    0.472   47.2           7.1
  1   0.428  0.028  0.215    1.090   60.0           7.4
  1   0.588  0.135  0.235    0.685   60.0           6.9
  1   0.438  0.805  0.222    0.573   60.0           6.5

     TIME REQUIRED TO OBTAIN THIS SOLUTION:    17 MIN
---------------------------------------------------------------------------------
```

Table 4.12: Summary log of most promising SOLVE solution showing all 10 of the anticipated selenium sites of the SV3CP dimer.

This solution was cause for great excitement for a number of reasons. Inspection of the selenium site co-ordinates showed that with the exception of one, selenium 2, proposed selenium locations were not laying along cell axes, a notorious source of spurious sites for reasons previously discussed. Peak heights for all the heavy atom sites in the cross-validation difference Fouriers were high i.e. above a value of 5. Also the Z score was above the accepted critical 20 mark at 26.33 and the phasing figure of merit was above 0.5 at 0.56.

The SOLVE output files, plus a file defining the SV3CP amino acid sequence, were passed straight to RESOLVE for selenium site location improvement, phase improvement and automated model fitting. After running RESOLVE it is important to check the summary (*Table 4.13*) of the phasing figure of merit (FOM) vs. resolution (a starting FOM calculated by SOLVE); when data has a FOM of less than 0.6 it is of no use in phasing.

RES	FOM	FOM-smoothed
22.12	0.62	0.75
16.52	0.53	0.75
12.48	0.64	0.75
9.77	0.64	0.75
8.26	0.74	0.74
7.15	0.78	0.74
6.23	0.77	0.73
5.58	0.80	0.73
5.09	0.77	0.72
4.71	0.73	0.71
4.38	0.71	0.71
4.11	0.70	0.70
3.88	0.71	0.70
3.66	0.72	0.69
3.44	0.70	0.68
3.25	0.65	0.67
3.08	0.62	0.66

Mean FOM 0.65

Table 4.13: Phasing figure of merit (FOM) with data resolution (RES) for the MAD SV3CP-MAPI dataset*. When the FOM is less than 0.60 diffraction data is not adequate for MAD phasing. Note greatest anomalous signal seen between 5–6 Å resolution enforcing that high resolution data is not necessary for phasing purposes. No FOM values are quoted above a resolution of 3.08 Å as the relatively weak anomalous signal results in a value of less than 0.6.*

Even more indicative of a good solution was the corrected figure of merit of phasing; above 0.7 is good, above 0.8 is very good. The corrected figure of merit should improve through each cycle of RESOLVE. Very encouragingly the final corrected figure of merit following RESOLVE was 0.80 for this solution. Absolute proof of a meaningful solution is, of course, visual inspection of the electron density map output from RESOLVE (*Figure 4.22*) for protein solvent boundary and features of protein secondary structure. A clear solvent boundary and evidence of alpha helices and beta sheets were quite apparent.

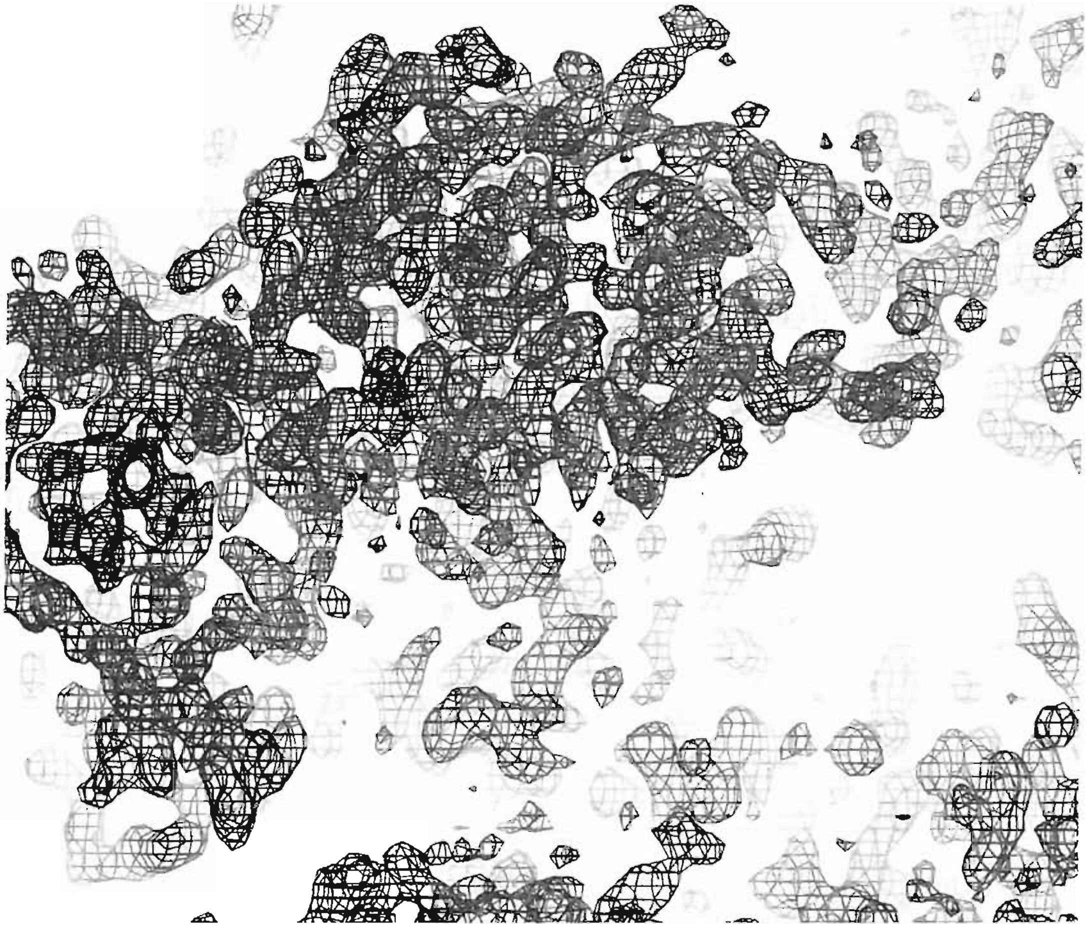


Figure 4.22: MAD phased electron density map of SV3CP-MAPI following phase estimate improvement with RESOLVE*. A clear protein solvent boundary and decipherable protein secondary structure are clearly visible. It should be noted these maps were produced prior to density modification in the CCP4 program: DM.*

Following the success of initial phasing it was disappointing to observe the effort made by RESOLVE at automated model building in attempting to fit the SV3CP peptide chain through the calculated electron density.

4.4.11 Density modification of MAD maps

Phases obtained from SOLVE and RESOLVE were further improved using the CCP4 program DM (density modification) (144) with the following procedures applied; solvent flattening, histogram matching, multi-resolution modification and non-crystallographic symmetry averaging. It should be noted that RESOLVE runs a similar but apparently less effective (judged by comparison of maps) routine as DM for phase improvement. Prior to DM, Matthews co-efficient (115) was run to ascertain crystal solvent content; calculating a co-efficient of 2.9 for 2 molecules in the asymmetric unit and a solvent content of 57.32%. Maps were generated using FFT (fast Fourier transform) (115) and inspected in the graphics package TURBO2000 and showed a considerable leap in quality and definition in comparison to those produced directly from RESOLVE.

4.4.12 Automated model building with MAID

Since automated model building was unsuccessful with RESOLVE, an alternative stand alone automated model building program; MAID (145) was used. The RESOLVE reflection file and SV3CP sequence was fed into MAID and on inspection of the resultant model a relatively impressive 60 % had been reasonably well fitted to the electron density (*Figure 4.23*). The remaining 40% of the model was either fitted incorrectly into density, fitted outside of density or completely absent. Notably poor regions were *N*-terminal residues Ala 1 to Thr 4, *C*-terminal residues Ser 174 to Glu 181; residues Leu 122 to Gly 140; a stretch that included the crucial active site Cys 139; and residues Lys 162 to Asp 165. Unfortunately, the version of MAID used offered no ligand fitting routine; therefore no attempt was made in fitting the MAPI. Further, it may well have been the presence of the peptidyl MAPI positioned in the SV3CP active site that contributed to the failure of MAID to fit this region, though density within this region was also very difficult to interpret as far as the appropriate density in which to fit protein and which to fit MAPI.

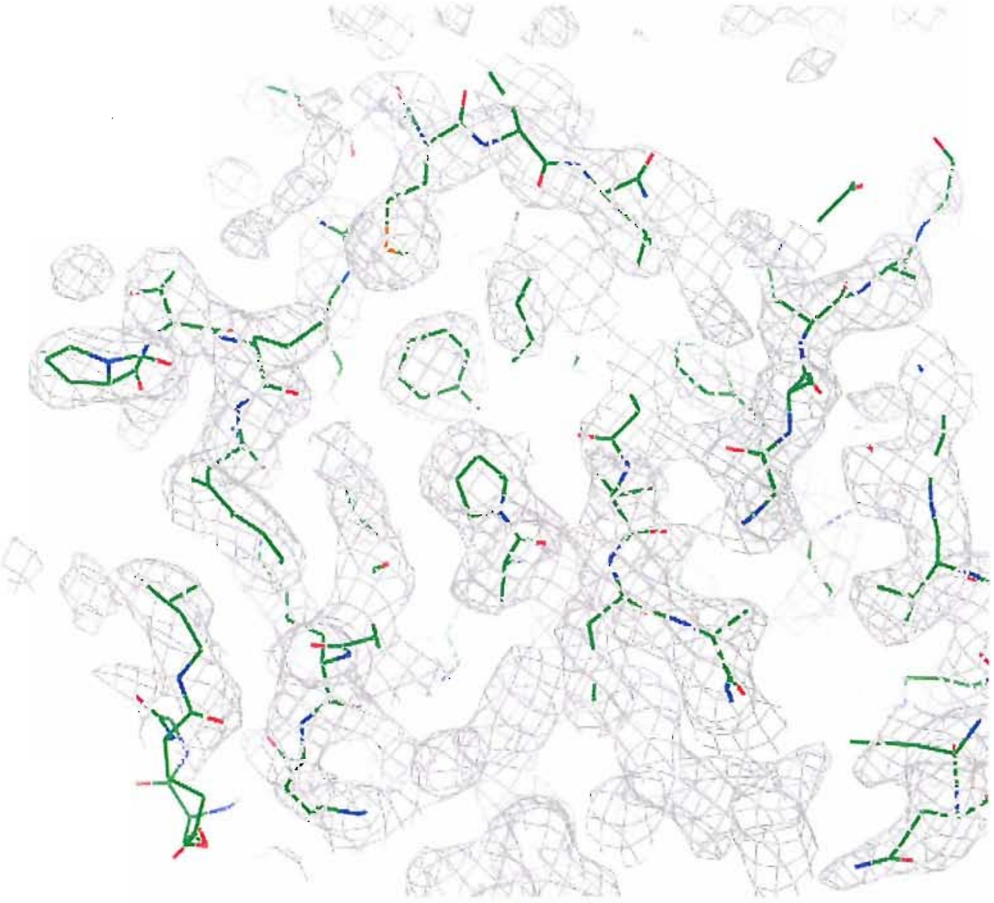


Figure 4.23: Partial model of SV3CP-MAPI produced by MAID shown with MAD phased electron density map to 2.2 Å resolution.

4.4.13 Manual model building

With a rough partial model obtained from MAID and quite easily interpretable maps, the first round of manual model building could begin. All model building was completed in TURBO2000. It should be noted that although the SV3CP was almost certainly as a dimer in the crystal, at this early stage model building was completed on a single SV3CP molecule only. This approach had two benefits. Firstly it required only one of the molecules of the dimer to be built to completion from the partial model provided by MAID. Once the first molecule had been completely built it was duplicated and moved as a fixed body to occupy the density that belonged to the second molecule of the dimer.

This model then required only relatively minor adjustment to accurately fit its density, therefore saving much time compared with separately building each molecule of the simer. Secondly building just one molecule at first was of benefit when deciding which density belonged to its dimer partner. It would be relatively easy to build a second molecule into the density belonging not to the dimer partner of the first molecule but into density belonging to a molecule related by crystallographic symmetry. However such an error can be avoided by building just one molecule of a suspected dimer then applying crystallographic symmetry operators that define your spacegroup to that molecule. It will become immediately apparent which density belongs to the dimer partner of your fully built molecule and which belongs to symmetry related molecules. This is all easily achievable using TURBO2000. On a separate issue, but of interest, regions of sketchy density in both molecules of the dimer highly corresponded to portions of the molecule that MAID had failed to fit and that therefore had to be built manually.

Following the first round of manual building, a model was achieved covering the majority of the SV3CP molecule with the exception of the MAPI bound to the active site CYS 139. Encouragingly density for the MAPI was well defined with both the side chains of the peptide portion and the *C*-terminal Michael acceptor extension instantly recognisable. At this early stage the MAPI was intentionally not built into density for ease of structural refinement. Protein regions where building had proved impossible were those already identified as having poor density; *N*-terminal residues Ala 1 to Thr 4, *C*-terminal residues Ser 174 to Glu 181, residues Leu 122 to Gly 140, and residues Lys 162 to Asp 165. Within TURBO2000 the option to display crystallographic symmetry related molecules was selected so that density belonging to the second molecule in the dimer could be identified (i.e. density without a corresponding model due to symmetry related molecules). As with molecule one of the dimer, secondary structural detail was quite apparent in the density of molecule two. It was then possible to locate the partial model manually built into molecule one of the dimer into the density for molecule two by treating the model as a rigid body.

4.5 Structural refinement

4.5.1 Refinement in SHELX

Using the peak dataset only, a free-R set was selected comprising of 5% of reflections using SHLEXPRO. The peak dataset and partial model now including co-ordinates for both molecules of the dimer, were passed to SHELXL (131) for a preliminary round of rigid body refinement, the result of which indicated an R-free of 37.28% and R-factor of 34.07%. Though both R values were relatively high this was as expected; it was the first round of refinement completed on a MAD structure with an incomplete model. Initial high R values are implicit in solving a novel structure. SHELXPRO was then used to calculate the sigmaA weighted $2F_o$-F_c and the F_o-F_c difference maps. Then followed two rounds of model building, and positional refinement with restraints in SHELXH with resultant refinement statistics of: R-free of 33.2% and an R factor of 27.23%. Unfortunately no further progress had been made with the problematic regions at this stage and still the MAPI had been left out for convenience (*Figure 4.24*).

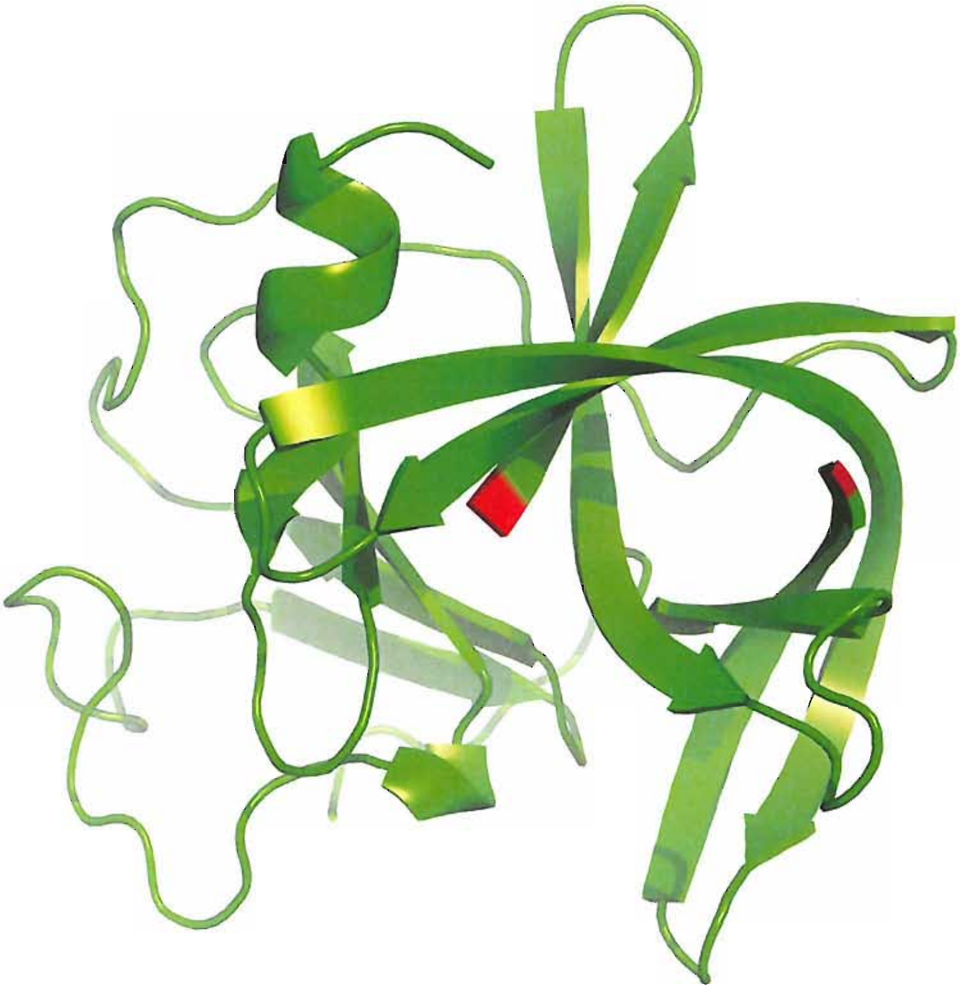


Figure 4.24: Early build model of SV3CP-MAPI*. The active site loop is missing, the beginning and end of which are indicated by residues in red. The MAPI had not been built in at this stage.*

4.5.2 Integration of a higher resolution native dataset

To allow building of the problematic regions and also so that a model based upon higher resolution data than the peak MAD dataset at 2.2 Å, the MAD peak data was discarded in favour of the previously collected 1.75 Å data of the native SV3CP in complex with the MAPI. A free-R set comprising 5% of native dataset reflections was selected using SHELXPRO. Without further model building a round of refinement in SHELXH followed by a round of model building and a further round of refinement resulting in reduced refinement statistics of: R-free of 30.64% and an R-factor of 27.1%. Immediate benefits of incorporating the higher resolution native data may not have been seen in largely improved refinement statistics, indeed that was not expected, but in the vast improvement in quality of maps. With features as detailed and informative as holes through tryptophan side chain rings and main chain carbonyl groups, model building was made far easier. Also density was much improved in the previously problematic regions and allowed these regions to be built with more confidence.

4.5.3 Refinement in CNS

From this point onwards the choice of structural refinement program was switched from SHELX to CNS (124) simply due to user preference. With the model in an improving state, it was time to build in the
MAPI. Since the MAPI is a novel compound, prior to the first round of refinement in CNS it was necessary to construct a ligand library manually; a list of all atoms and bonds within the MAPI along with bond length and angles used for refinement purposes. For the peptide portion this was a straight-forward case of looking up ideal values for the relevant amino acid. For the MAPI functional group; the *C*-terminal propenyl-ethyl-ester extension comprising the Michael acceptor, bond lengths and angles had to be manually calculated through a combination of rudimentary chemical principles and basic trigonometry, as did values defining the covalent bond to the active site Cys 139 side chain sulphur. It was the relative complexity of writing a ligand library and defining non-peptide bonds for refinement purposes that the MAPI had to this point been omitted from refinement. On technical issues; refinement was completed in CNS MINIMISE, for positional refinement, then CNS B-INDIVIDUAL, for temperature factor refinement, with sigmaA weighted $2F_o$-F_c and F_o-F_c maps calculated in CNS MODEL MAP, and converted into TURBO format with the UPPSALLA program MAPMAN. By round 4 of refinement, model fit and the associated phase improvement was sufficient to identify and build water molecules into stereo-chemically acceptable positions. Water molecules were

placed in appropriate positive density of the F_o-F_c difference map contoured at 3 σ that were within 3.5 Å of the protein surface and formed hydrogen bonds with chemically acceptable groups. Following a subsequent round of refinement, the $2F_o$-F_c maps were inspected to confirm water molecules remained within density.

As a final solution was approached, alternate side chain conformations were built in for residues; Gln 172 of molecule one, and Ser 7, Thr 28, Val 85, Met 120, Met 130, of molecule two. Throughout refinement the region spanning Asn 126 to Leu 132 of molecule 1 suffered poorly defined density. It was not so much due to an absence of density, more the density was broad and without easily discernable features to assist building. Ultimately, this region was built with the Cα peptide backbone in two entirely separate conformations (*Figure 4.24*), and, after much manipulation, this region refined satisfactorily. It was apparent that this region was quite flexible; explanation is offered in *Section 5.1.5*. Interestingly, the same region in molecule two showed far more easily interpretable density and only a single conformation of the Cα peptide backbone could be fitted albeit in a different conformation to either of those in molecule 1.

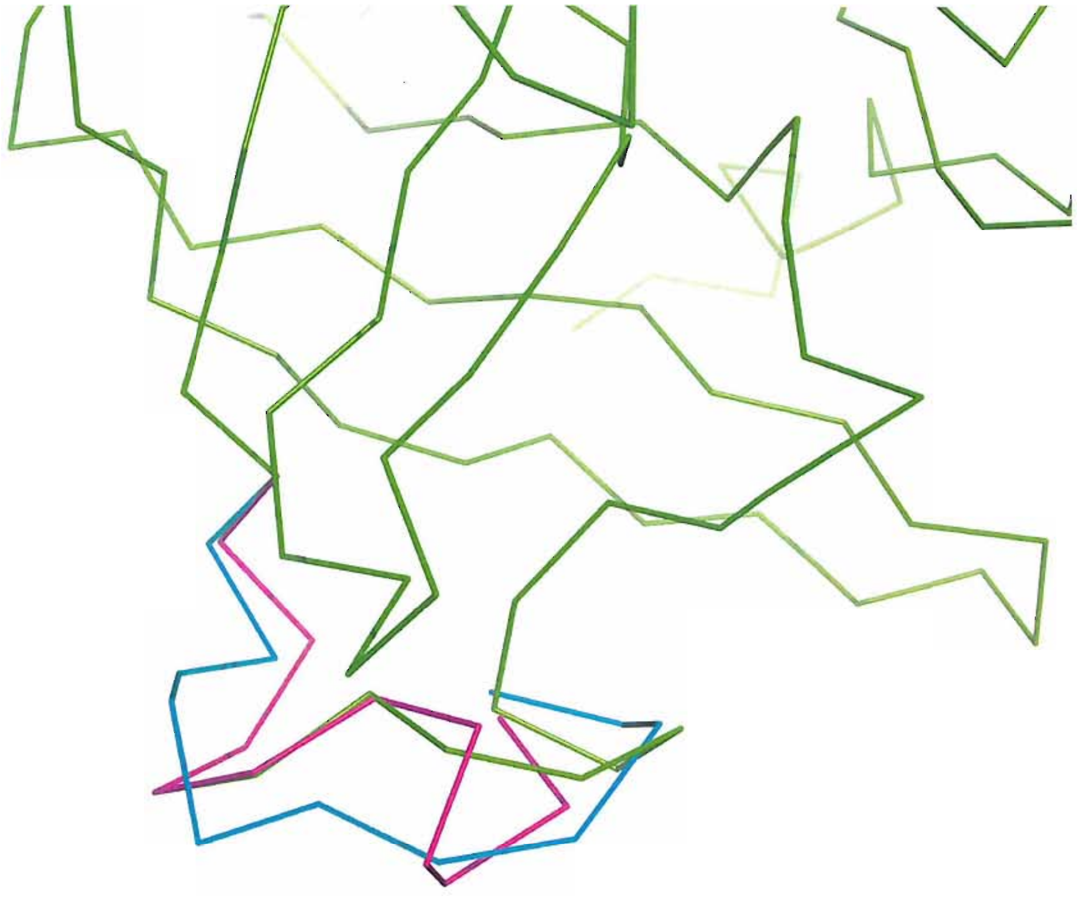


Figure 4.25: The three conformations of the Asn 126 to Leu 132 loop as seen in the SV3CP-MAPI dimer*. The conformations shown in green and pink are those built as alternate conformations into the density of molecule 1 of the dimer. The loop shown in blue shows a third conformation of the active site loop as seen in molecule 2 of the dimer; molecules 1 and 2 have been aligned to best show this.*

In total, 22 rounds of model building and refinement were completed in CNS. Residues that proved impossible to fit convincingly into density included; the final *C*-terminal residue Glu 181 of both molecule one and molecule two, and the *N*-terminal residues Ala 1, Pro 2 and Pro 3 of molecule 2. Missing *N*- and *C*-terminal residues are relatively common in crystallographic structures. Final refinement statistics were; an overall R_{free} of 22.27% and an R factor of 20.48% (*Table 4.14*). Data quality was also reflected in the appearance and features of the final maps (*Figure 4.25 and 4.26*); only rarely was side chain density partial; long side chains such as lysines and arginines showed well defined density; all main chain carbonyl "bumps" could be seen and holes through aromatic side chains were apparent, in one case through the side chain of Pro 22. In addition to the protein model 180 water molecules were identified within a 3.5 Å limit of the protein surface.

Refinement range / Å	100 – 1.75
R_{factor}	20.48
R_{free}	22.27
Total number of reflections in resolution range	51760 (99.6 %)
Number of reflections in working set	49180 (94.6 %)
Number of reflections in test set	2580 (5 %)
Number of protein atoms	2870
Number of solvent atoms	255

Table 4.14: Final refinement statistics for the structure of SV3CP in complex with MAPI.

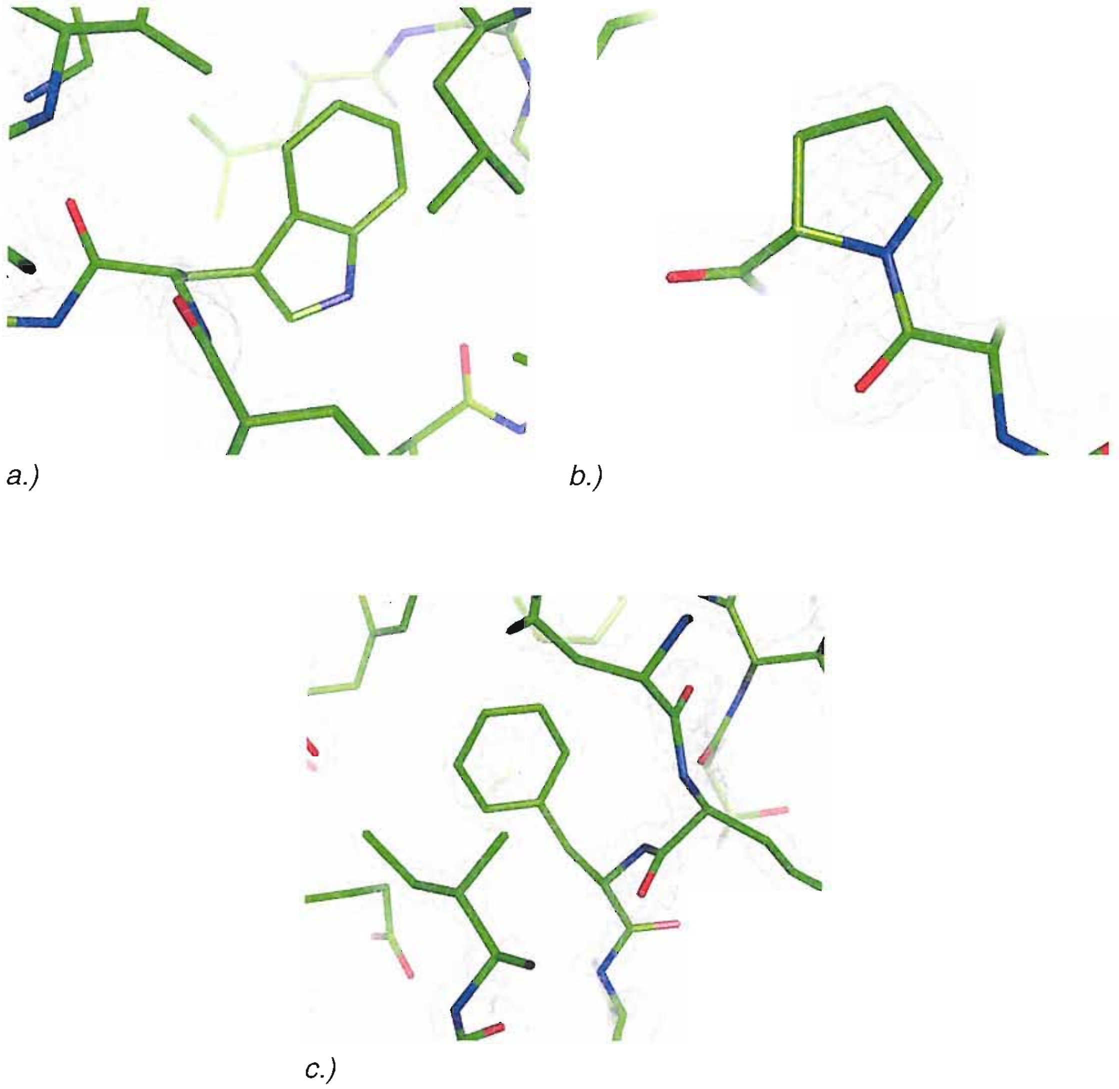


Figure 4.26: $2F_o$-F_c sigmaA-weighted electron density map of SV3CP-MAPI. *The high qaulity of map is reflected in holes through density of a.)Trp 16, b.) Pro 33, and c.) Phe 58 side chains. Maps show 1.75 Å resolution data.*

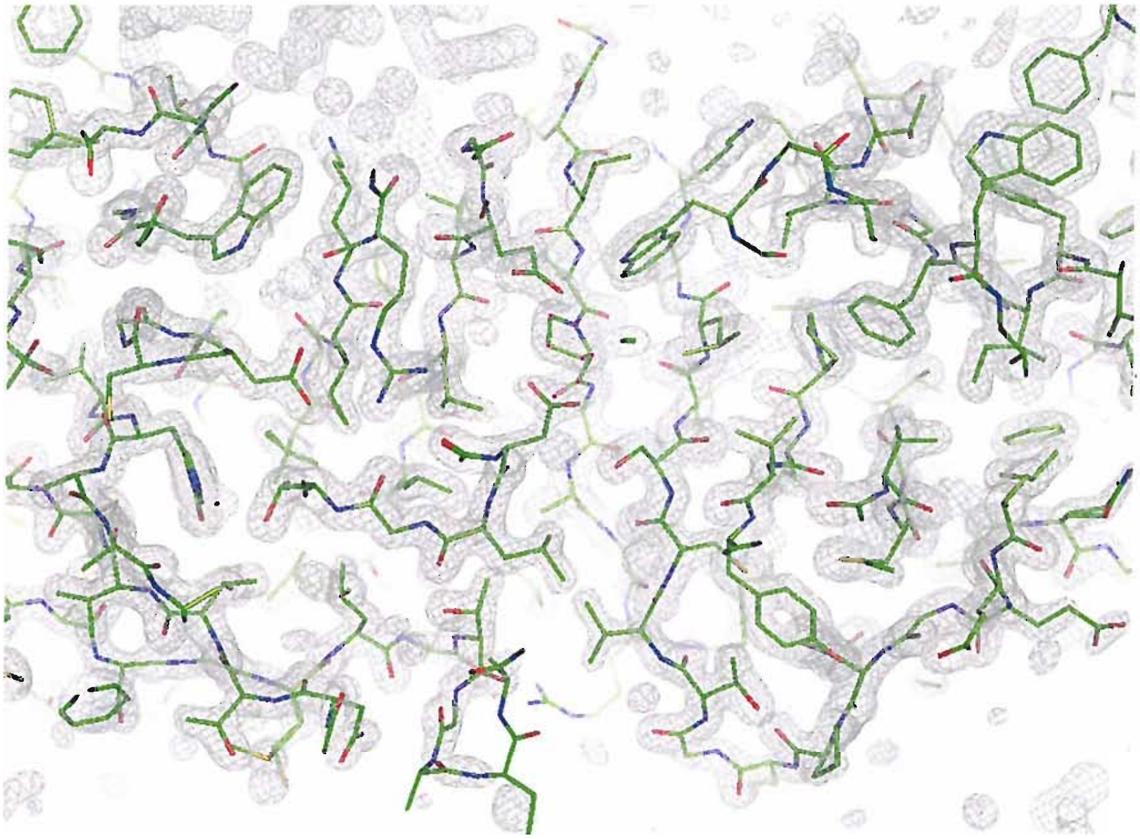


***Figure 4.27: Demonstration of overall quality of electron density**. $2F_o$-F_c sigmaA-weighted electron density map calculated from 1.75 Å resolution data.*

4.5.4 Technical overview of a structural solution of SV3CP-MAPI

Outlined in *Figure 4.28* is a summary of the major computational steps taken in solving the structure of SV3CP in complex with MAPI by MAD phasing. Accompanying vital statistics are included where appropriate.

MOSFLM

To autoindex and integrate images

Cell parameters: a = 49.4973 b = 84.1056 c = 121.4705 α = β = γ = 90.000

Space group: P $2_12_12_1$

↓

SORTMTZ (CCP4)

To arrange mtz file for SCALA

↓

SCALA (CCP4)

To scale and merge (native datasets only) reflections.

Native data set; d_{min} = 1.75 Å, R_{merge} = 0.053; MAD datasets: Peak: d_{min} = 2.26 Å, R_{merge} = 0.086; Inflection point: d_{min} = 2.5 Å, R_{merge} = 0.057; Remote: d_{min} = 2.5 Å, R_{merge} = 0.064.

↓

TRUNCATE (CCP4)

To convert structure factor intensities into structure factor amplitudes

↓

MTZUTILS (CCP4)

To create rearranged dataset from multiple datasets, i.e. for MAD data to merge three datasets into one.

↓

SCALEIT (CCP4)

To scale a derivative dataset to a native dataset i.e. for MAD data to scale multiple wavelengths usually to the peak data.

↓

SOLVE

To determine ASA sites, derive ASA phases, extrapolate ASA phases to protein atoms.*

**ASA-anomalously scattering atoms – i.e. Seleniums*

Z score = 26.33 , Phasing FOM = 0.56.

↓

RESOLVE

To refine ASA positions and build peptide sequence into density.

Stating phasing FOM = 0.65; finishing FOM = 0.80

(automated build function not used for SV3CP-MAPI complex)

↓

DM – Density Modification (CCP4)

To improve protein – solvent boundary of density.

↓

FFT – Fast Fourier Transform

To produce electron density maps from structure factors.

Within this thesis to produce sigmaA weighted 2Fo-Fc and Fo-Fc difference maps.

↓

MAID – Automated Model Building

Attempts to best fit peptide sequence of protein to density.

Fitted 60% of SV3CP though only quite approximately.

↓

TURBO2000

Graphics package.

Used in electron density inspection and manual model building.

↓

SHELX

Positional refinement with restraints .

Within this thesis, completed 2 rounds of refinement before switching to CNS for refinement;

R =27.1, R_{free} = 30.64

↓

CNS – Generate, Alternate, Minimise, B-individual

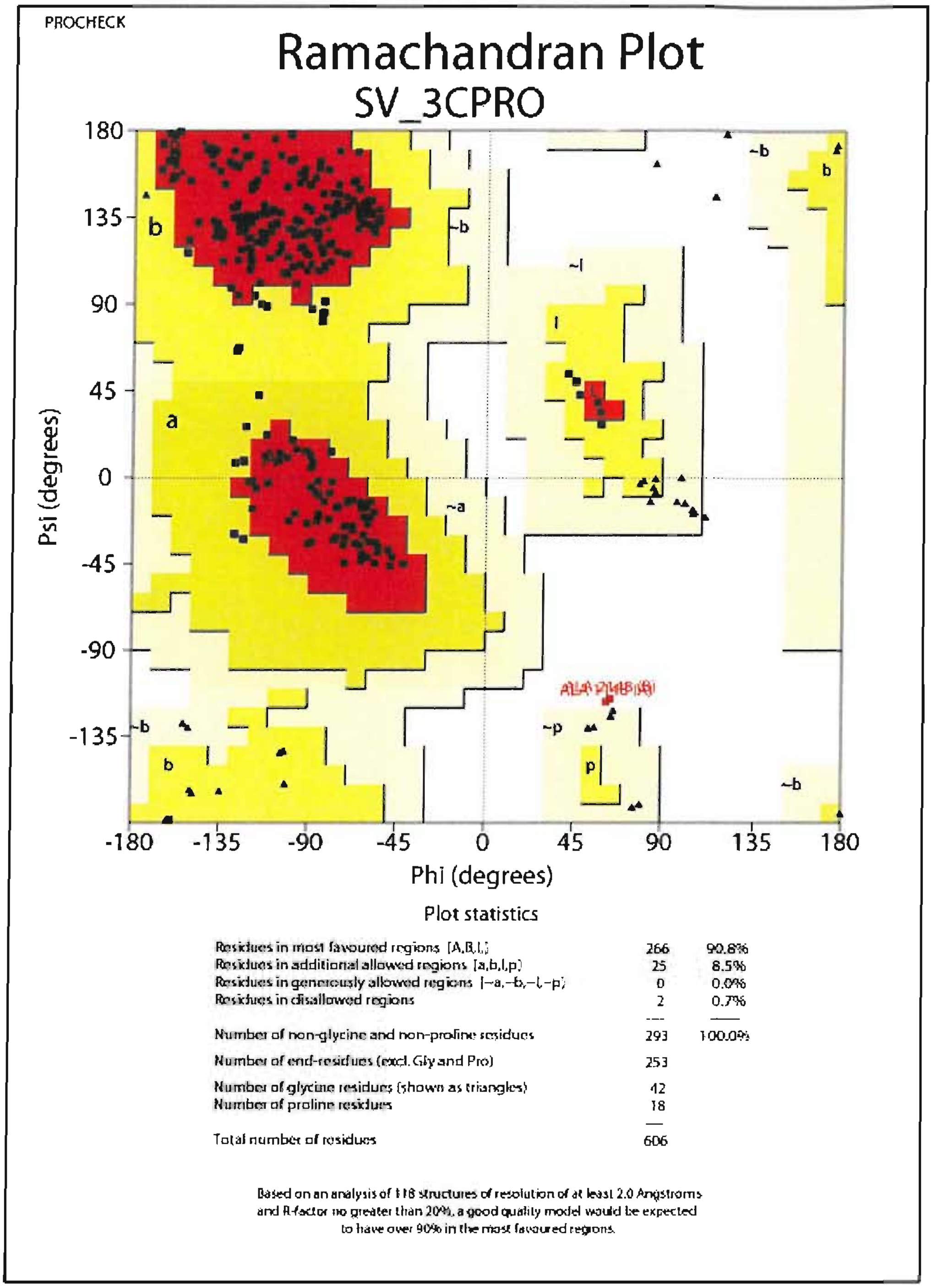
Refinement suite; Generate: to produce CNS format files; Alternate; to define alternate conformations of individual residues. Minimise: atomic positional refinement with restraints; B-Individual: independent atomic temperature factor refinement.

22 rounds of refinement in total; final R = 20.46, R_{free} = 22.27

Figure 4.28: Overview (accompanied by appropriate statistics for SV3CP-MAPI complex) of the computational steps taken in solving the X-ray crystallographic structure of SV3CP in complex with MAPI.

4.6 Structural validation

Protein structure stereo-chemical qualities were assessed using the CCP4 suite program; PROCHECK (146) and independently using MOLPROBITY (147) of the RCSB server. Results vary between validation programs. By PROCHECK Ramachandran analysis (*Figure 4.28*), 90.8% of residues were shown in the most favoured regions, 8.5% in allowed regions with 0.7% in disallowed regions. The two residues in the disallowed regions were Ala 149 from both molecule 1 and molecule 2 of the SV3CP dimer. Ala 149 is found at the apex of a loop region connecting anti-parallel beta sheets (for full structure description see *Chapter 6*). Repeated attempts were made to fit Ala 149 into alternate conformations to satisfy Ramachandran analysis but ultimately the best fit to density of Ala 149 and of neighboring residues was when in the disallowed conformation, possibly due its stressed position at the apex of a hairpin loop. Separate analysis by MOLPROBITY places all residues within allowed regions of the Ramachandran plot (*Figure 4.29*); with 98.3% in favored regions and 1.7% in permissible regions. Therefore a discrepancy exists over whether Ala 149 is in an allowed or disallowed conformation dependent upon whether structural analysis is completed by PROCHECK and by MOLPROBITY.


PROCHECK

Ramachandran Plot

SV_3CPRO

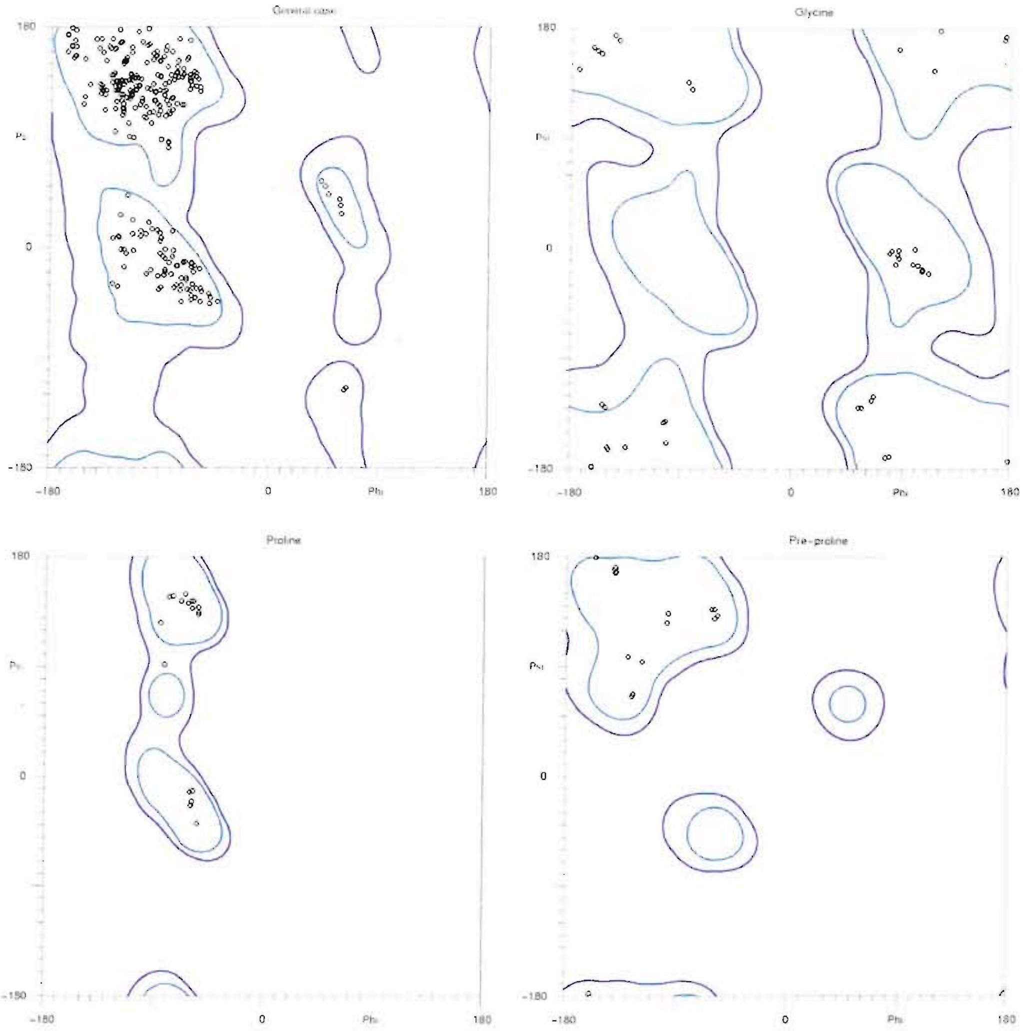
Plot statistics

Residues in most favoured regions [A,B,L]	266	90.8%
Residues in additional allowed regions [a,b,l,p]	25	8.5%
Residues in generously allowed regions [~a,~b,~l,~p]	0	0.0%
Residues in disallowed regions	2	0.7%
	----	----
Number of non-glycine and non-proline residues	293	100.0%
Number of end-residues (excl. Gly and Pro)	253	
Number of glycine residues (shown as triangles)	42	
Number of proline residues	18	

Total number of residues	606	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

Figure 4.29: Ramachandran plot of the SV3CP-MAPI structure following final refinement. *Plot produced according to PROCHECK structural analysis.*



353 residues were evaluated in total for general, glycine, proline, and pre-pro
98.30% of all residues were in favored (98%) regions (347 residues)
100.00% of all residues were in allowed (>99.8%) regions. (353 residues)

There were no outliers

Lovell, *et al.* (2003) Proteins: Struct Func Gen **50**:437-450

Figure 4.30: Ramachandran plot of the SV3CP-MAPI structure following final refinement*. Plot produced according to MOLPROBITY structural analysis.*

Further measurement of stereo chemical quality is offered by main chain and side chain bond length and angle analysis in PROCHECK (*Table 4.15*). Average temperature factor can also be indicative of a fair model; 27.27 is quite an acceptable value.

Parameter	RMSD from ideal values
Covalent bond lengths (Å)	0.005
Covalent bond angles (°)	1.3
Average B-factor (Å^2)	27.27

Table 4.15: Stereo chemical parameter values as determined by the structural validation program PROCHECK.

The structural co-ordinates have been submitted along with structure factor amplitudes (and phases) to the RCSB Protein Data Bank and can be found under accession number: *2IPH at www.rcsb.org.*

5.0 Discussion

5.1 General structural features of SV3CP

The high quality and resolution of the final electron density map allowed for the positioning of main-chain, side-chain, MAPI and 180 solvent atoms with confidence. Commensurate with crystallographic structures of other viral 3C proteases (poliovirus (164), HAV (167, 168) and HRP (165) the SV3CP structure displays the same chymotrypsin-like fold comprising of two domains connected by a large loop, typical of this group of proteases (31). Domain I consists of a short N-terminal two turn α-helical portion leading directly into the first of five β-strands: βaI to βeI (see *Table 5.1* and *Figures 5.2 to 5.5*), that comprise a twisted anti-parallel β-sheet. Connected to domain I via a 20 residue loop; domain II consists of six β-strands: βaII to βfII that comprise an anti-parallel β-barrel (see *Figures 5.2* to *5.5*). The primary amino acid sequence of SV3CP with associated topology is shown in *Figure 5.1*.

Domain	Secondary structure	Identifier	Residue range	Length of region (aa's)
I	α-helix	αaI	Pro3-Arg8	5
	β-strand	βaI	Val9-Lys11	3
	loop	lpaI	Phe12-Gly15	4
	β-strand	βbI	Trp16-Ser21	6
	loop	lpbI	Pro22,Thr23	2
	β-strand	βcI	Val24-Thr28	5
	loop	lpcI	Thr29-Ile47	19
	β-strand	βdI	Ala48-Ala52	5
	loop	lpdI	Gly53,Glu54	2
	β-strand	βeI	Phe55-Arg59	5
I/II (inter-domain loop)	loop	lpI/II	Phe60-Glu79	20
II	β-strand	βaII	Gly80-Arg89	10
	loop	lpaII	Asp90,Ser91	2
	β-strand	βbII	Gly92-Ile109	18
	loop	lpbII	Gln110,Gly111	2
	β-strand	βcII	Arg112-Leu121	10
	loop	lpcII	Leu122-Cys139	18
	β-strand	βdII	Gly140-Arg147	8
	loop	lpdII	Ala148,Asn149	2
	β-strand	βeII	Asp150-Ala160	11
	loop	lpeII	Thr161-Asn165	5
	β-strand	βfII	Thr166-Ala170	5
	loop	lpfII	Val171-Leu180	10

Table 5.1: Secondary structural features of SV3CP.

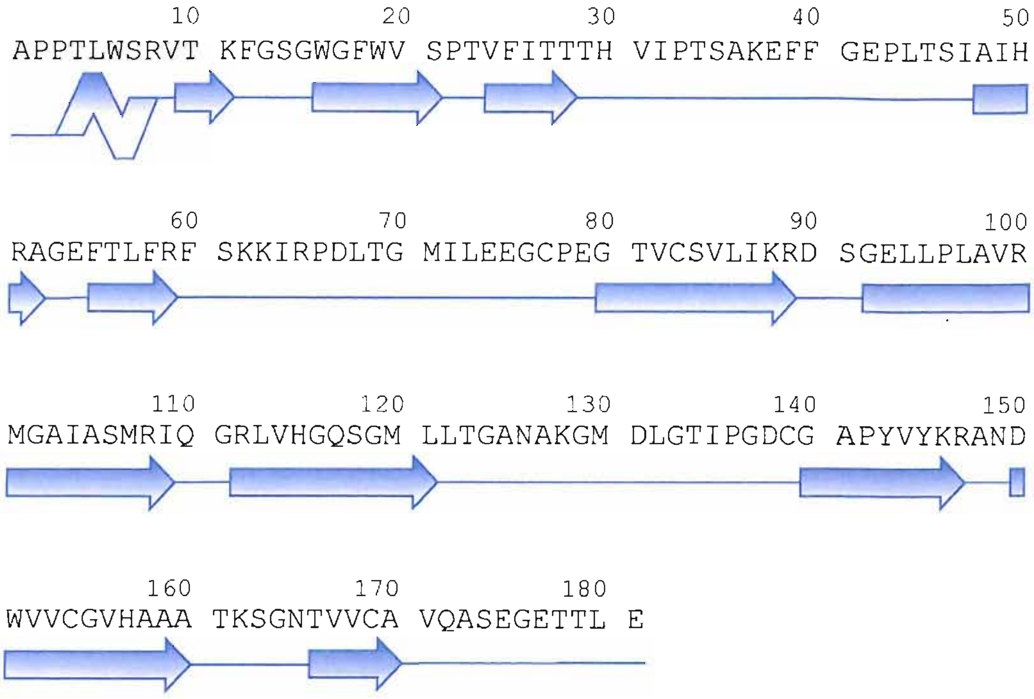

10 20 30 40 50
APPTLWSRVT KFGSGWGFWV SPTVFITTTH VIPTSAKEFF GEPLTSIAIH
60 70 80 90 100
RAGEFTLFRF SKKIRPDLTG MILEEGCPEG TVCSVLIKRD SGELLPLAVR
110 120 130 140 150
MGAIASMRIQ GRLVHGQSGM LLTGANAKGM DLGTIPGDCG APYVYKRAND
160 170 180
WVVCGVHAAA TKSGNTVVCA VQASEGETTL E

Figure 5.1: The primary amino acid sequence of SV3CP*. The associated topology is represented as 'zig-zags' for alpha helices and 'arrows' for β-strands.*

A cursory inspection of the SV3CP-MAPI structure indicates that Cys139 is the catalytic cysteine since the Cys139 sulfhydryl can clearly be seen to be covalently bound to the Michael acceptor (MA) extension of MAPI (see *Figure 5.5*). In addition, hitherto ambiguity as to whether the SV3CP possesses a catalytic diad or triad of residues is partially resolved as two further members of the catalytic domain: Glu 54 and His 30 are within plausible bonding distances as to complete a triad of catalytic residues (see *Figures 5.3* and *5.4*). A E54A mutant (169) has been proved to diminish protease activity, strongly implicating it plays a role in substrate catalysis. Both glutamate and histidine are common members of catalytic triads of cysteine proteases.

Also of note is the orientation of the β-strands: βbII and βcII, as they not only contribute to the β-barrel of domain II but also form an "arch" through which the substrate binding S sites coordinate the peptide portion of MAPI (see *Figure 5.5*). A SV3CP structure without substrate bound (or peptide based inhibitor as presented here) may wrongly suggest that substrate would bind through the cleft generated between domains I and II.

It becomes clear that the β-barrel structural motif of domain II does not act simply as a scaffold to co-ordinate the active site catalytic residues but is directly involved in binding substrate.

Further dissection of the SV3CP catalytic domain and coordination of MAPI within the active site cleft is offered later in this chapter.

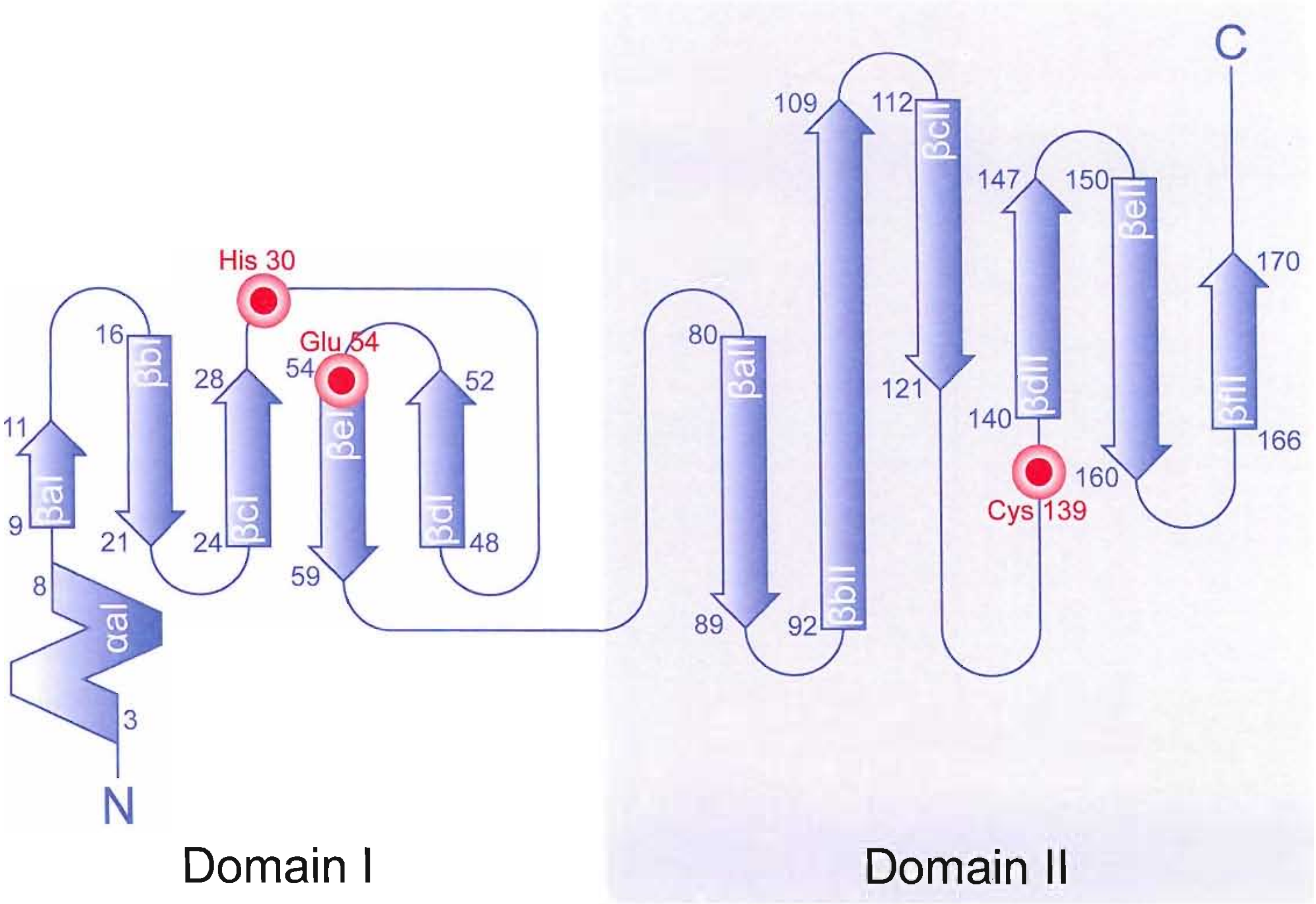


Figure 5.2: Secondary structure topology diagram of SV3CP*. Active site catalytic residues: Cys139, Glu 54 and His 30 are indicated in red. With SV3CP correctly folded the catalytic residues are brought within acceptable bonding distances to form the active site catalytic triad.*

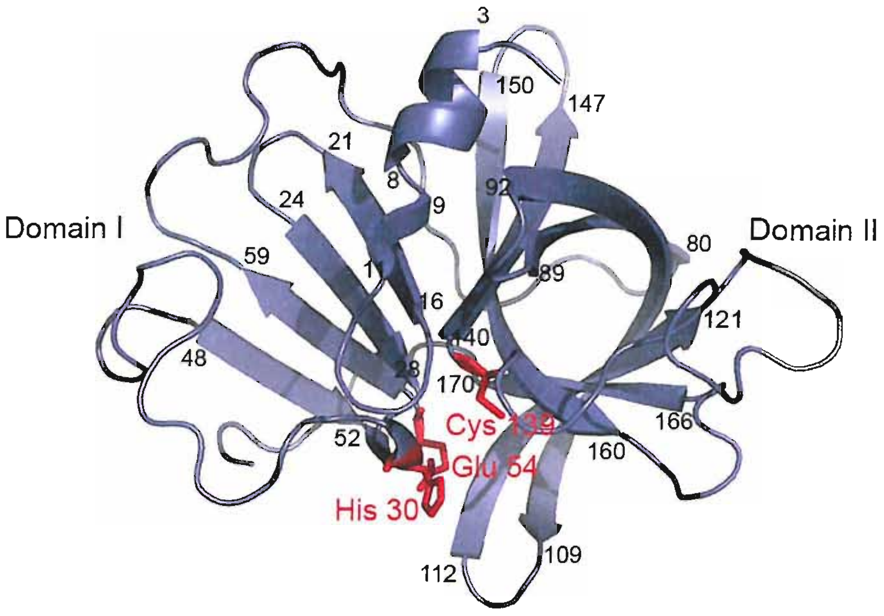


Figure 5.3: SV3CP viewed through the β-barrel of domain II from the N-terminal side. *The active site catalytic triad of residues: Cys 139, His 30 and Glu 54 are indicated in red. MAPI is not shown. Other residues are indicated numerically only.*

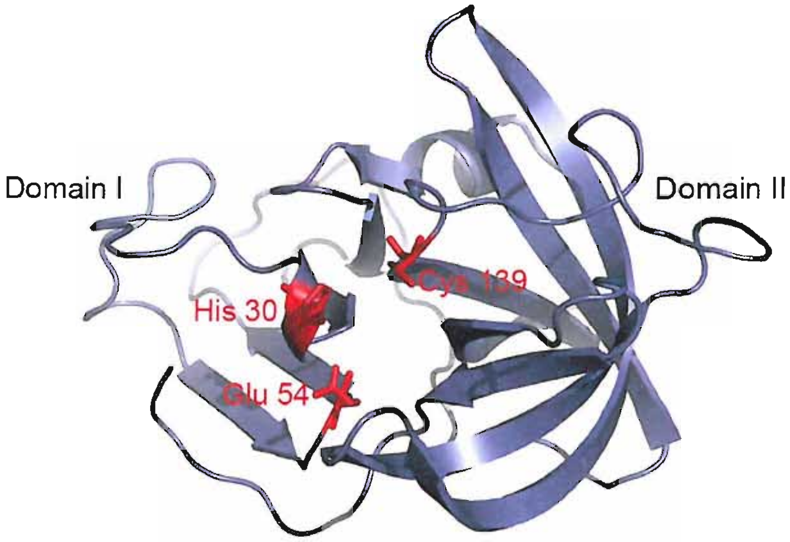


Figure 5.4: SV3CP rotated 90° from Figure 5.3. *The close proximity of the three members of the catalytic triad is more easily appreciated. MAPI is not shown.*

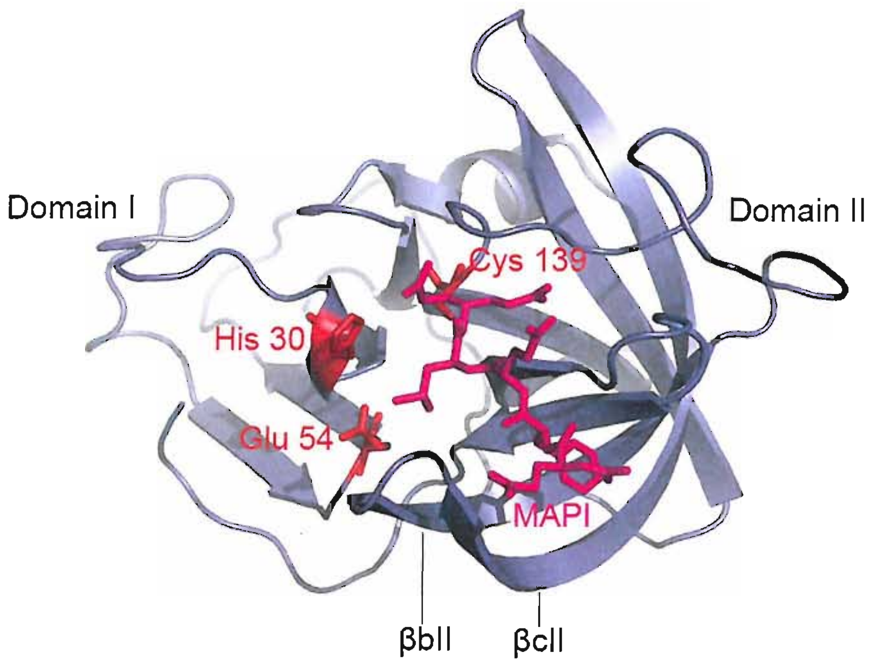


***Figure 5.5: SV3CP as Figure 5.4 but with MAPI shown**. Observe how β-strands βbII and βcII create an arch to coordinate the peptidyl portion of MAPI (corresponding to the natural substrate P residues).*

5.1.1 Domain I of SV3CP displays an incomplete β-barrel motif

The twisted β-sheet motif of domain I reflects a tertiary structural difference to the poliovirus, HAV and HRP 3C-proteases that display a complete β-barrel in this domain. The long loop: lpcI, consisting of 19 residues (Thr 29-Ile 47) connecting βcI to βdI, lacks any distinguishable secondary structure and it seems plausible that if this portion were to possess a more β-strand like characteristic, domain I may form a complete β-barrel conformation. Further evidence that domain I can be considered an incomplete β-barrel rather than a twisted β-sheet is offered in the form of a hydrophobic core, a common feature of β-barrel motifs. The domain I hydrophobic core is comprised of the hydrophobic residues: Phe 12, Trp 19, Phe 25, Ile 32, Phe 39, Phe 40, Ile 47, Ile 49, and Ile 64 (see *Figure 5.6*).

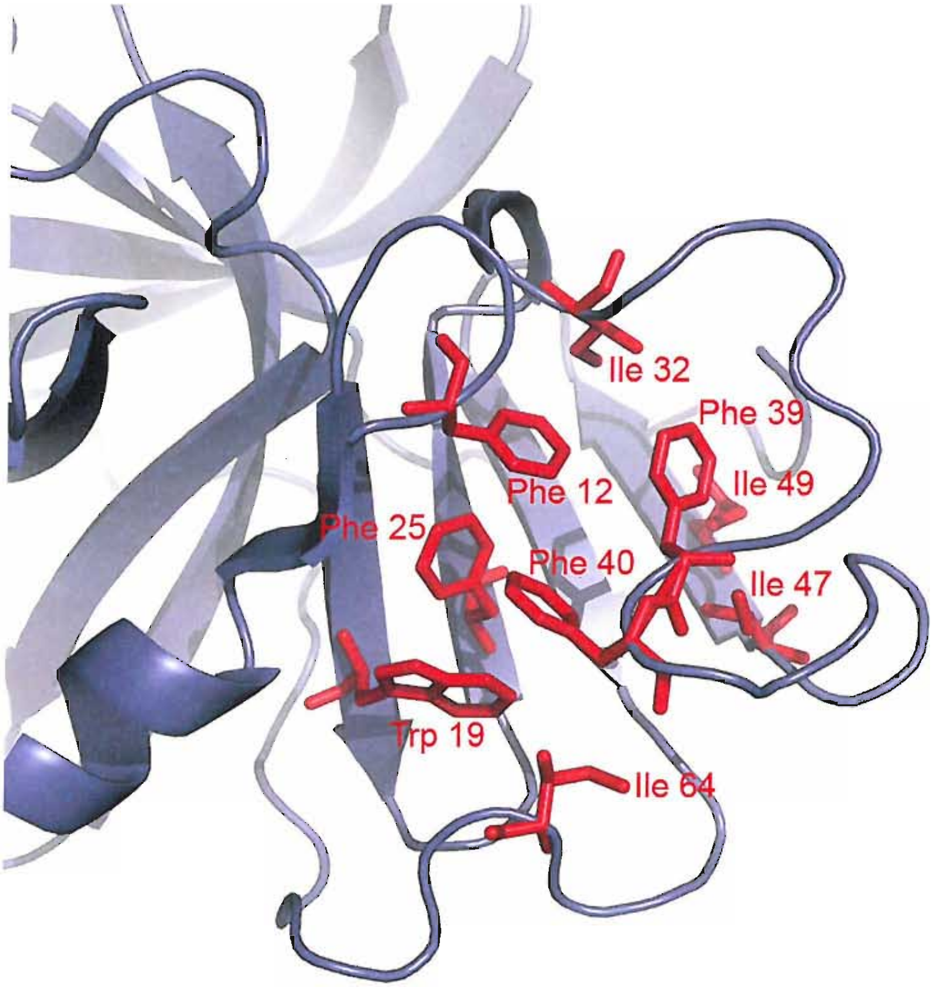


***Figure 5.6: The hydrophobic core of the incomplete β-barrel of domain I**. The hydrophobic core is comprised of Phe 12, Trp 19, Phe 25, Ile 32, Phe 39, Phe 40, Ile 47, Ile 49 and Ile 64, all are shown in red.*

5.1.2 Domain II displays a complete β-barrel motif

In contrast to domain I, domain II displays a complete β-barrel comprised of the key residues: Val 85, Val I99, Leu 121, Val 167, and the less hydrophobic, but contributory residues: Tyr 143 and His 157 (see *Figure 5.7*). This β-barrel motif will later be shown to be important in the orientation of the residues Tyr 143 and His 157 that are intrinsic to substrate binding at the S1 site.

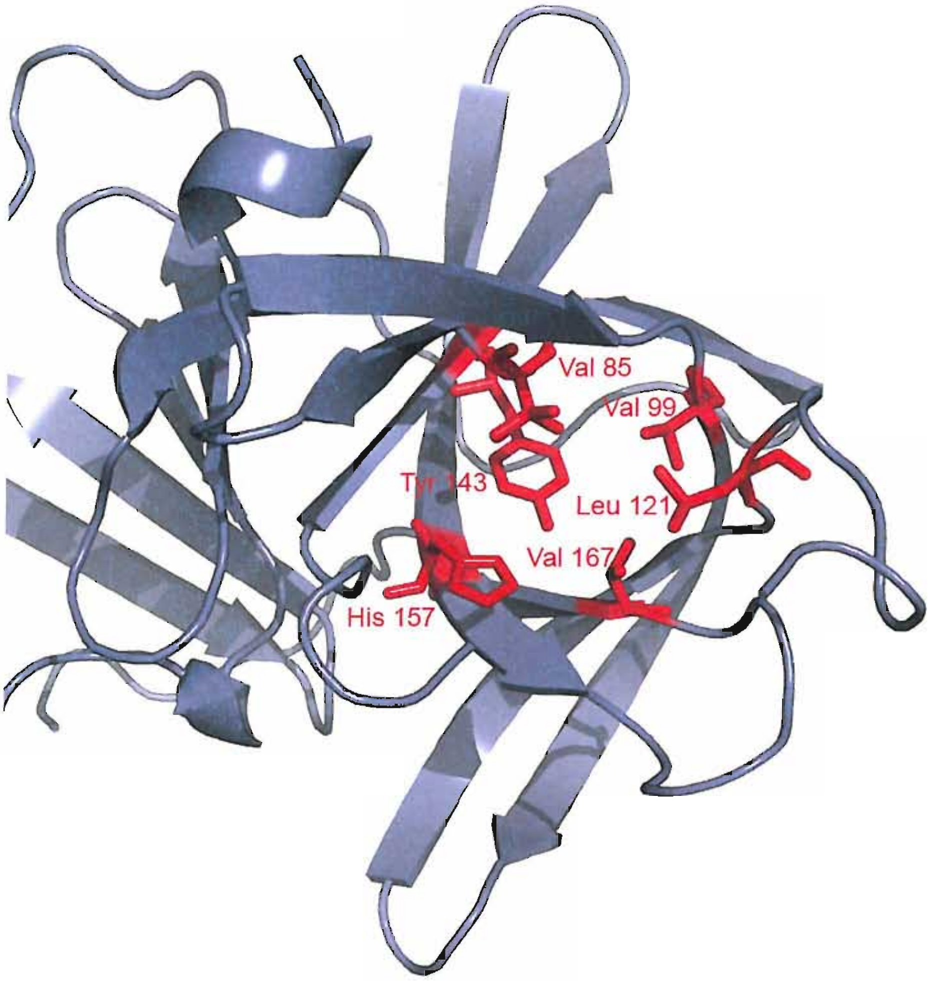


Figure 5.7: The hydrophobic core of domain II. *The hydrophobic core is comprised of the residues: Val 85, Val 99, Leu 121, Tyr 143, His 157 and Val 167.*

5.1.3 SV3CP dimerises

Analysis by application of the Matthews co-efficient (62.3% solvent, co-efficient of 2.89) prior to obtaining a final structural solution indicated the presence of two molecules of SV3CP in the asymmetric unit.

The dimer interface (see *Figure 5.8*) is maintained by a network of hydrogen bonding between residues of strands: αaI, lpcII, βaII and βbII, resulting in a total buried surface area of 2,353 Å^2 calculated using ArealMol (115). All intermolecular electrostatic interactions are listed in *Table 5.2*.

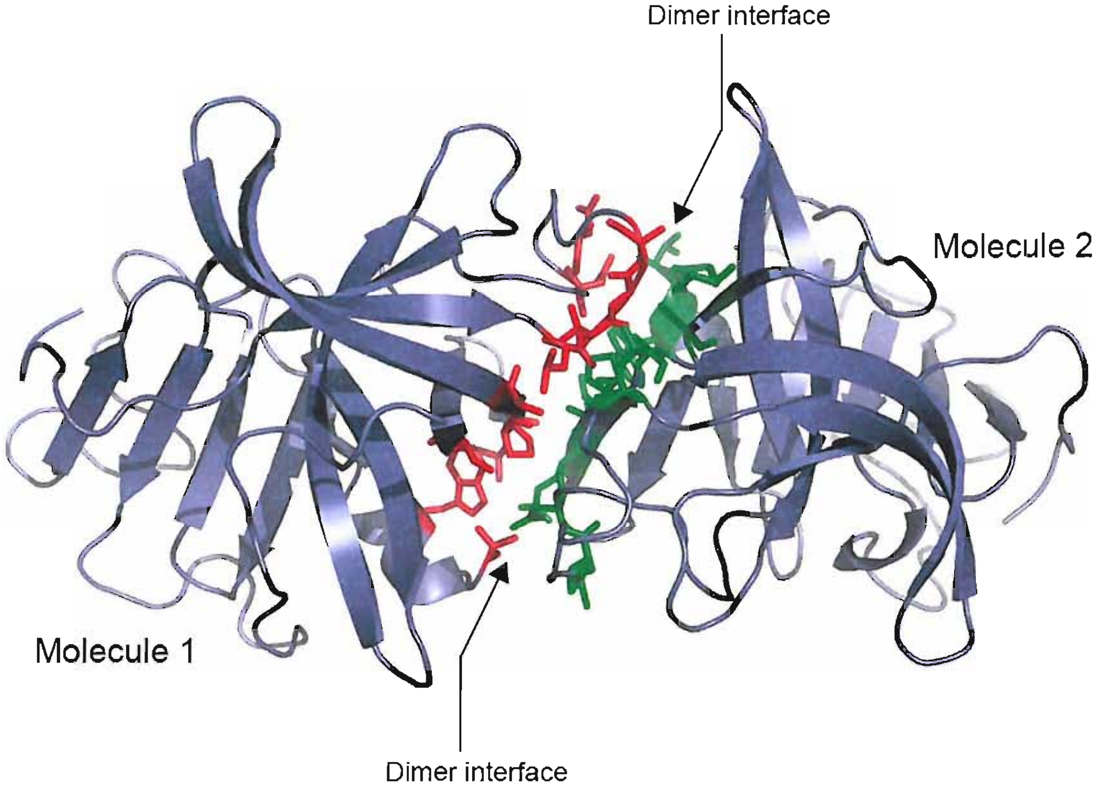


***Figure 5.8: Dimeric arrangement of SV3CP**. Residues involved in dimerisation are shown in red for molecule 1 and green for molecule 2. MAPI is not shown.*

Molecule 1	Molecule 2	Interaction via water molecule / or direct	Bonding type	Bond length / Å
Ala1 N	Asp131 Oδ1 and Oδ2	direct	H-bond	2.88 and 2.68
Ala1 N	Glu93 Oε2	direct	H-bond	2.80
Ala1 N	Met130 Sδ	via water molecule	H-bond	3.03
Trp6 NE1	Glu93 Oε2	direct	H-bond	2.75
Ser84 Oγ	Asp131 Oδ2	via water molecule	H-bond	2.78 (2.89)
Ser84 Oγ	Asp131 O	via water molecule	H-bond	2.71 (2.70)
Leu94 O	Leu94 N	direct	H-bond	2.76
Pro96 O	Pro96 O	via water molecule	H-bond	3.04 (2.73)
Pro96 O	Asp131 O	via water molecule	H-bond	2.72 (2.70)
Ala 98 N	Pro96 O	via water molecule	H-bond	2.87 (2.73)
Arg100 NH1 and NH2 (of Cζ)	Leu122 O	direct	H-bond	3.03 and 2.70
Leu122 O	Ala98 N	direct	H-bond	2.85
Leu122 O	Ala98 O	via water molecule	H-bond	3.10 (2.75)
Leu122 O	Leu122 N	via water molecule	H-bond	3.10 (3.09)
Thr123 O	Ser84 Oγ	direct	H-bond	2.62
Ala125 N	Val82 O	via water molecule	H-bond	2.91 (2.77)
Asp131 O	Trp6 Nε1	direct	H-bond	3.16
Asp131 Oδ2	Leu5 N	via water molecule	H-bond	2.58 (3.14)
Asp131 Oδ2	Thr4 N	via water molecule	H-bond	2.58 (2.82)

Table 5.2: Electrostatic interactions that maintain the dimer interface between molecules 1 and 2 of the SV3CP dimer. *When an electrostatic interaction by hydrogen bonding is made via an intermediate water molecule the distance from water molecule to the relevant residue of molecule 2 is quoted in parenthesis. Direct hydrogen bonds are highlighted in blue.*

Many of the residues involved in maintaining the dimer interface interact through hydrogen bonding via bridging water molecules. Though these interactions may stabilise the SV3CP dimer, the more important electrostatic interactions are direct hydrogen bonds either between main-chain atoms of separate residues (*main-chain to main-chain* interactions), or a main-chain atom of one residue to a side-chain atom of a second (*main-chain to side-chain* interactions). *Main-chain to main-chain* hydrogen bonding exists between the main-chain O of Leu 94 of molecule 1 and main-chain N of Leu 94 of molecule 2, and separately, between the main-chain O of Leu 122 and main-chain N of Ala 98. These intermolecular interactions are likely to be highly important in maintaining the SV3CP dimer. Additionally the *main-chain to side-chain* interactions between: the main-chain N of Ala 1 of molecule 1 and Oδ2 of Asp 131 of molecule 2; both NH1 and NH2 of Cζ belonging to Arg 100 of molecule 1 and the main-chain O of Leu 122 of molecule 2; the main-chain O of Thr 123 of molecule 1 and the Oγ of Ser 84 of molecules 2; and the main-chain O of Arg 131 of molecule 1 and the Nε1 of Trp 6 of molecule 2, also all appear to be of importance in maintaining the dimer interface. Finally, a single *side-chain to side-chain* hydrogen bond exists between Nε1 of Trp 6 of molecule 1 and Oε2 of Glu 93 of molecule 2 to further stabilise the SV3CP dimer. All residues involved in the dimer interface with the exception of Ser 84, which is substituted for Thr in some *Noro*viral strains, are highly conserved within the *Noro*virus family 3C proteases suggesting that dimerisation may be required for structural integrity and presumeably protease activity.

The SV3CP dimer interface presented here is at odds with observations made by Zeitler (2006) (169) of the 3C-protease belonging to the Norwalk *Noro*viral strain. Zeitler suggests the residues: Glu 79, Ser 106, Arg 108, Ser 163, Thr 166, Glu 177 and Glu 181 are involved in intermolecular dimer interactions; however, these residues are located on an entirely different face of the molecule to those involved in the SV3CP dimer interface. This begs the question as to which of the dimer interfaces is correct; that proposed for SV3CP or that for the Norwalk virus 3C protease, or alternatively, whether the dimer interface differs between *Noro*viral strains? In answer to the later suggestion, sequence and structural comparison suggest it would be unlikely that two so closely related proteases should have evolved different modes of dimerisation. In answer to which structure presents the correct mode of dimerisation it may be helpful to consider the nature of crystallisation of each protease. As already mentioned the asymmetric unit of the SV3CP-MAPI crystals contains two molecules of SV3CP arranged in a dimer i.e. this dimeric arrangement is not as a result of the crystal lattice. In contrast, the Norwalk-virus structure has one molecule in the asymmetric unit therefore interactions between molecules may be due to crystal contacts and not representative of the formation of the natural dimer. It may be argued therefore that a higher level of confidence can be placed in observations pertaining to dimerisation of the SV3CP than the Norwalk-3C-protease structure.

5.1.4 Are both molecules of the SV3CP dimer functionally active?

Three-dimensional superimposition of molecules 1 and 2 (see *Figure 5.9*) of the dimer reveals an r.m.s. deviation of just 0.414 Å. The r.m.s. deviation would be even lower at 0.292 Å with the flexible region, Thr 123 to Asp 131 (see *Section 5.1.5*), omitted. This almost insignificant deviation of main-chain atoms may indicate that both molecules of the dimer are functionally active. Since the SV3CP has been proven active by colorimetric assay (see S*ection 4.2.4*), the single conformation displayed by both subunits of the dimer must be an active conformation. In addition, both molecules of the dimer have MAPI bound, further evidence that as a dimer SV3CP may bind and correctly orientate substrate (the peptide portion of MAPI being analogous to substrate) in a functional binding cleft and active site. The opposite might be true if a large r.m.s. deviation were seen between the two molecules of the dimer since it would be unlikely that two protease molecules adopting significantly different conformations would both be functionally active. Though as already discussed this is not the case with the SV3CP dimer.

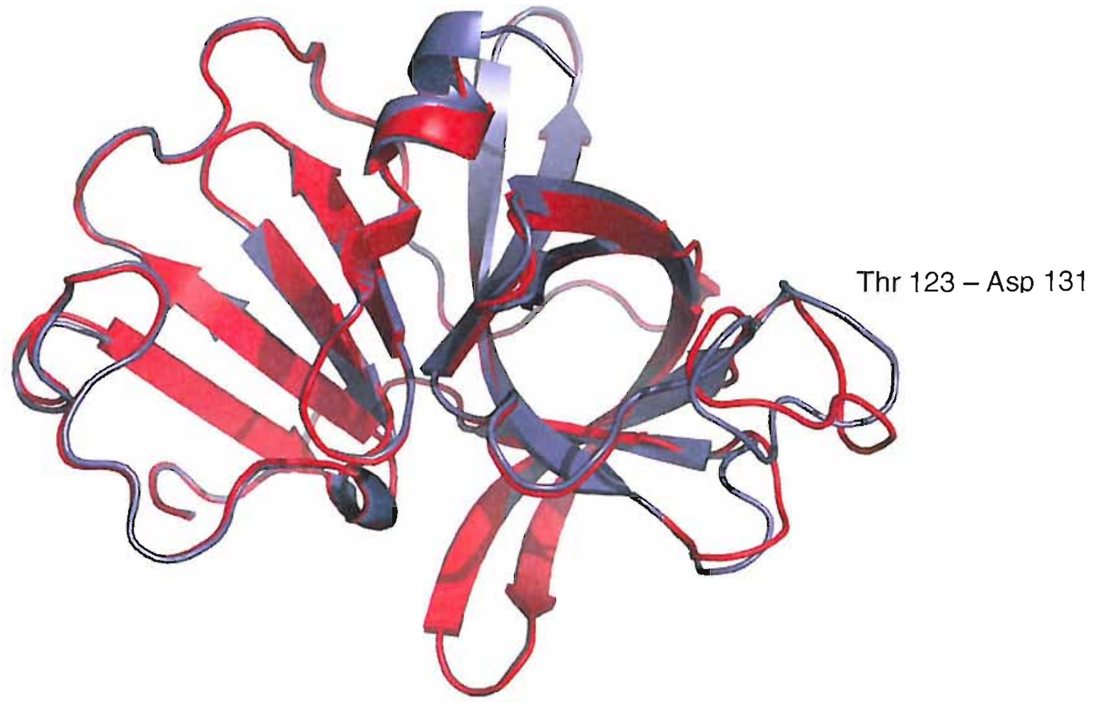
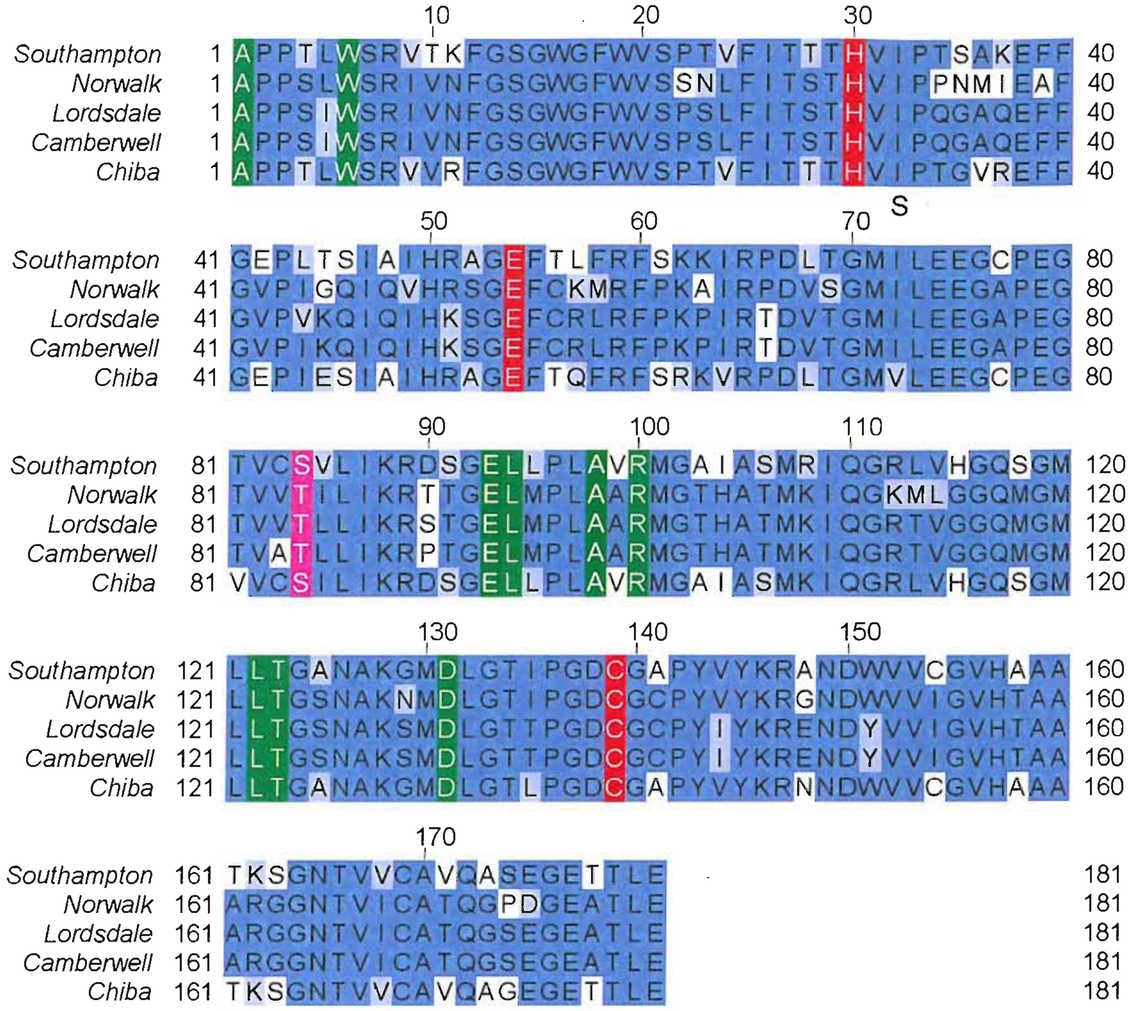


***Figure 5.9: Superimposition of molecule 1 (blue) and molecule 2 (red) of the SV3CP dimer**. Molecules 1 and 2 of the SV3CP dimer adopt almost identical conformations. The exception to this is the flexible loop region comprised of the residues Thr 123 to Asp 131.*



```
                           10        20        30
Southampton   1 APPTLWSRVTKFGSGWGFWVSPTVFITTTHVIPTSAKEFF 40
    Norwalk   1 APPSLWSRIVNFGSGWGFWVSSNLFITSTHVIPPNMIEAF 40
  Lordsdale   1 APPSIWSRIVNFGSGWGFWVSPSLFITSTHVIPQGAQEFF 40
 Camberwell   1 APPSIWSRIVNFGSGWGFWVSPSLFITSTHVIPQGAQEFF 40
      Chiba   1 APPTLWSRVVRFGSGWGFWVSPTVFITTTHVIPTGVREFF 40
                                                 S
                           50        60        70
Southampton  41 GEPLTSIAIHRAGEFTLFRFSKKIRPDLTGMILEEGCPEG 80
    Norwalk  41 GVPIGQIQVHRSGEFCKMRFPKAIRPDVSGMILEEGAPEG 80
  Lordsdale  41 GVPVKQIQIHKSGEFCRLRFPKPIRTDVTGMILEEGAPEG 80
 Camberwell  41 GVPIKQIQIHKSGEFCRLRFPKPIRTDVTGMILEEGAPEG 80
      Chiba  41 GEPIESIAIHRAGEFTQFRFSRKVRPDLTGMVLEEGCPEG 80

                           90       100       110
Southampton  81 TVCSVLIKRDSGELLPLAVRMGAIASMRIQGRLVHGQSGM 120
    Norwalk  81 TVVTILIKRTTGELMPLAARMGTHATMKIQGKMLGGQMGM 120
  Lordsdale  81 TVVTLLIKRSTGELMPLAARMGTHATMKIQGRTVGGQMGM 120
 Camberwell  81 TVATLLIKRPTGELMPLAARMGTHATMKIQGRTVGGQMGM 120
      Chiba  81 VVCSILIKRDSGELLPLAVRMGAIASMKIQGRLVHGQSGM 120

                          130       140       150
Southampton 121 LLTGANAKGMDLGTIPGDCGAPYVYKRANDWVVCGVHAAA 160
    Norwalk 121 LLTGSNAKNMDLGTIPGDCGCPYVYKRGNDWVVIGVHTAA 160
  Lordsdale 121 LLTGSNAKSMDLGTTPGDCGCPYIYKRENDYVVIGVHTAA 160
 Camberwell 121 LLTGSNAKSMDLGTTPGDCGCPYIYKRENDYVVIGVHTAA 160
      Chiba 121 LLTGANAKGMDLGTLPGDCGAPYVYKRNNDWVVCGVHAAA 160

                          170
Southampton 161 TKSGNTVVCAVQASEGETTLE                   181
    Norwalk 161 ARGGNTVICATQGPDGEATLE                   181
  Lordsdale 161 ARGGNTVICATQGSEGEATLE                   181
 Camberwell 161 ARGGNTVICATQGSEGEATLE                   181
      Chiba 161 TKSGNTVVCAVQAGEGETTLE                   181
```

***Figure 5.10: Multiple sequence alignment of the Noro*viral family *3C proteases*.** *Sequence alignment was performed using CLUSTALX. Sequence comparison was completed using JALVIEW with the BLOSUM62 scoring matrix employed to score sets of aligned residues; a deeper blue indicates a more highly conserved sequence. The residues His 30, Glu 54 and Cys 139 that comprise the active site catalytic triad are highlighted in red. Conserved residues involved in maintaining the dimer interface; Ala 1, Trp 6, Glu 93, Leu 94, Ala 98, Arg 100, Leu 122, Thr 123 and Asp 131 are highlighted in green; the variable residue Ser/Thr 84 is highlighted in pink. The sequences were obtained from the UniProtKB/SWISSPROT sequence databank with the accession numbers: Southampton virus Q04544; Norwalk virus A0ZNP2; Lordsdale virus P54634; Camberwell virus Q9W183; and Chiba virus Q9DU47.*

5.1.5 The presence of a flexible loop of domain II close to the active site catalytic triad

Within domain II of SV3CP exists a long flexible loop region comprised of the residues Thr123 to Asp131. This flexible region was refined into two conformations within molecule 1 of the dimer and a third conformation within molecule 2 (see *Section 4.3.5* and *Figure 4.24*). Lying close to the active site catalytic triad it is quite possible this disordered loop may interact with substrate P' residues and so contribute to substrate binding. Unfortunately any specifics of such interactions cannot be reported here since MAPI does not include residues C-terminal to the scissile bond to mimic the natural substrate P' residues. With such residues not present in the SV3CP-MAPI structure, whether this flexible loop interacts with substrate P' residues remains unclear. However, it follows that a loop that would normally bind substrate may become disordered in the absence of substrate on account of stabilising loop-to-substrate interactions being absent. The Norwalk (169) and Chiba (170) 3C protease structures lack any bound substrate and both show the expected flexibility in this region, supporting the theory this region may interact with substrate.

However, the proposal that this flexible loop interacts with substrate is just one possibility. An alternative and possibly more likely explanation may be that to meet intermolecular interactions between the two molecules of the dimer a certain amount of flexibility in this region is required. From inspection of the dimer interface (see *Section 5.1.3*) it can be seen that Thr 123, Ala 125 and Asp 131 are all involved in intermolecular dimer interactions. This region must therefore be considered to play an important role in dimerisation. Interestingly Thr 123, Ala 125 and Asp 131 all interact with residues of the dimer partner molecule of SV3CP that are positioned within regions possessing a secondary structure: Thr 123 interacts directly with Ser 84 that lies within the βaII β-strand, Ala 125 interacts via a water molecule to Val 82 that also lies within βaII, and, Asp 131 interacts directly with Trp 6 and via water molecules to Thr 4 and Leu 5 that all lay within the N terminal α-helix: αaI. Regions possessing secondary structure are *relatively* inflexible, and as such, if involved in intermolecular interactions as here, may require the region containing the residues with which they interact to display a degree of flexibility in order to bring interacting residues into sufficient proximity to electrostatically interact. Such a hypothesis may account for the flexibility of the region Thr 123 to Asp 131 in both molecules of the dimer.

5.1.6 A cysteine cluster in domain II

On first inspection, domain II contains what appears to be a 'cysteine cage', a rather rare structural feature, comprising of the residues Cys 77, Cys 83, Cys 154 and Cys 169 (see *Figure 5.11*). A cysteine cage normally consists of four cysteine residues coordinated by a centrally located metal ion with distances between the sulphurs of each cysteine of approximately 3.5 Å. Though distances between cysteine sulphurs in the SV3CP structure are slightly greater than this, in the absence of a coordinating metal ion some flexibility leading to greater distances may be expected. Indeed, on inspection of the 2Fo-Fc and Fo-Fc density maps there exists a complete lack of density at the centre of the cysteine cluster indicating the absence of a coordinating metal ion. These observations cast doubt over whether this cluster of cysteines is as a true cysteine cage. Certainly, during kinetic analysis, the SV3CP showed activity without the need to supply an exogenous source of metal ions. Further, the lack of density appropriable to a metal ion rules out the proposal that a coordinating ion being constitutively bound during expression in *E.coli*. It is therefore concluded that SV3CP does not require to co-ordinate a metal ion for activity despite it possessing what appears to be a cysteine cage. As a result it seems more appropriate to call it a 'cysteine cluster' rather than a cysteine cage.

The real function of the cysteine cluster appears to be quite conventional. Sitting at the opposite end of the domain II β-barrel to the active site, the SV3CP cysteine cluster simply contributes to the hydrophobic core of the barrel as the sulphurs are likely to be protonated and interact hydrophobically. Strengthening the hydrophobic nature of this cluster is a further sulphur containing hydrophobic residue: Met101. Interestingly, the Chiba virus 3C protease includes this cysteine/methionine cluster whereas the Norwalk virus 3C protease does not.

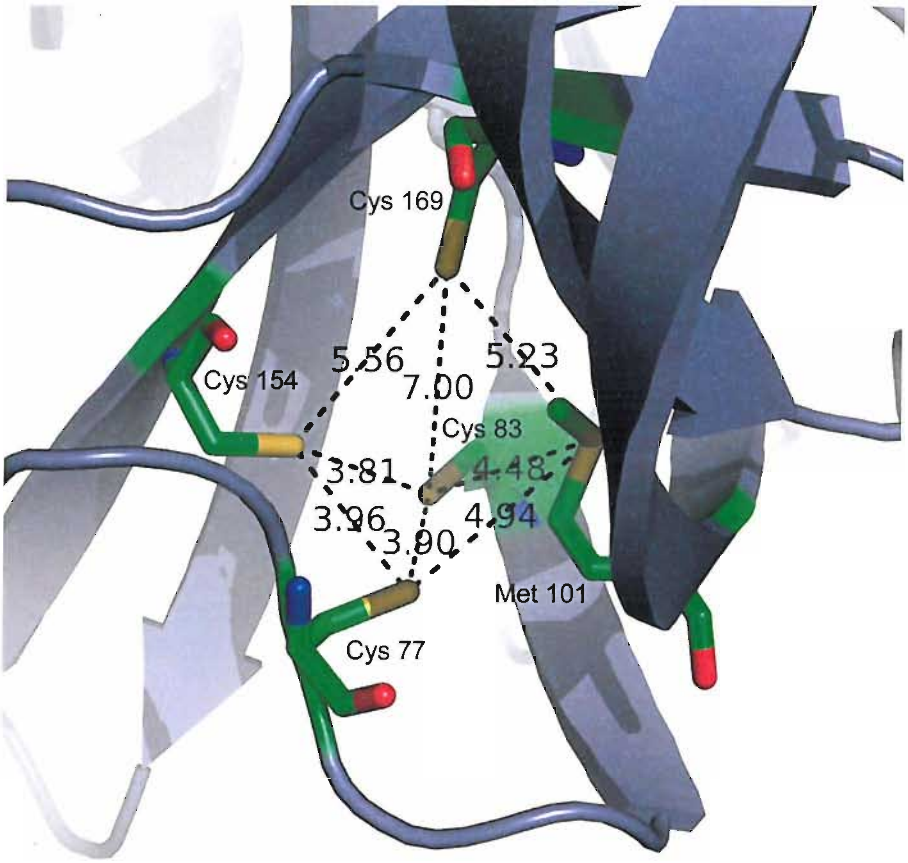


Figure 5.11: The cysteine/methionine cluster of the β-barrel of domain II *.The cysteine cluster of SV3CP contributes to the hydrophobic nature of the β-barrel of domain II.*

5.2 Binding of the peptidyl portion of MAPI by SV3CP

For ease of identification, the residues of the peptide portion of MAPI will be named: E5, F4, Q3, L2 and Q1 corresponding to the residues of MAPI in the N to C terminal direction: Glu, Phe, Gln, Leu and Gln respectively.

5.2.1 Limited interactions may be important for coordination of the P5 substrate residue

Whilst no definable SV3CP S5 pocket exists, limited interactions between MAPI E5 and SV3CP (see *Figure 5.12*) support preliminary kinetics work that suggests coordination of substrate at an S5 site may be important for substrate cleavage. The main-chain carbonyl O of MAPI E5 hydrogen bonds via a surface water molecule to the main-chain amine N of Lys162. Whilst not as robust as direct *main-chain to main-chain* interactions; *main-chain to main-chain* interactions via bridging water molecule are often of importance for enzyme-substrate coordination.

A second SV3CP-MAPI E5 interaction exists via a water molecule between the Oδ1 of E5 and the main-chain N of Gln 110. This interaction is unlikely to be as important for substrate coordination as the *main-chain to main-chain* interaction and almost certainly does not contribute to substrate sequence recognition by SV3CP despite the involvement of the MAPI E5 side-chain. Comparison of residues at the S5 position of the five points of cleavage within the natural substrate: Asp, Glu, Ser, Lys, Glu, reveals characteristically diverse residues. The apparent ability of SV3CP to coordinate such a range of residues at the substrate P5 position is presumably attributable to lack of a fully developed S5 binding site.

In summary, although SV3CP does coordinate substrate at an S5 substrate binding site, the identity of the substrate residue does not seem of importance, though the coordination of the substrate peptide backbone by SV3CP may well be. These observations support those of the preliminary kinetics work that show the penta-peptide chromogenic peptide experiences a greater rate of cleavage than the tetra-peptide chromogenic peptide.

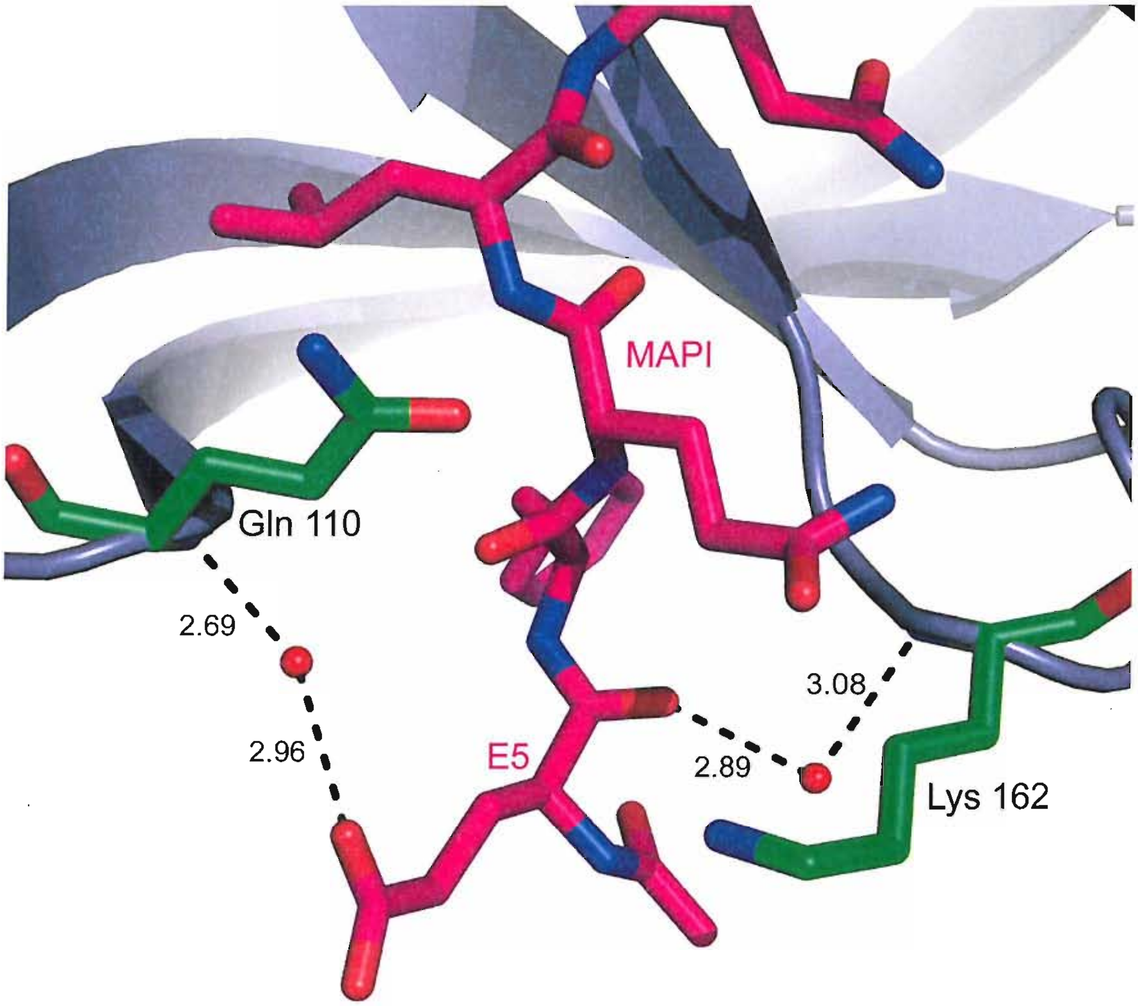


Figure 5.12: The limited electrostatic interactions between SV3CP and E5 of MAPI. *Water molecules shown as red spheres.*

5.2.2 A surface hydrophobic S4 binding pocket

The hydrophobic S4 binding pocket is comprised of the residues: Met 107, Ile 109, Thr 161, Thr 166 and Val 168 (see *Figure 5.13*). The hydrophobicity of the S4 binding pocket is underlined by a complete absence of solvent molecules. The S4 pocket lies on the surface of SV3CP and as such may be considered a hydrophobic *surface* pocket. Cleavage sequence comparison of the five ORF1 substrate recognition sites at the substrate P4 position: Phe, Phe, Ala, Ile and Thr, reveals residues that are either highly hydrophobic: Phe and Ile, or slightly hydrophobic: Ala and Thr. The nature of the pocket itself, and the residues at the substrate P4 position, indicates very strongly the need for SV3CP to accommodate a hydrophobic residue within the S4 binding pocket.

In addition to hydrophobic interactions of the MAPI S4 side-chain with SV3CP S4 pocket residues, the SV3CP-MAPI structure shows interactions between main-chain MAPI and main-chain SV3CP atoms. The MAPI F4 main-chain amine N and carbonyl O interact by hydrogen bonding via separate ordered water molecules to the carbonyl O of Arg 108 and main-chain amine N of Gln 110 respectively. Though these interactions do not confer substrate specificity, as they only involve atoms of the substrate peptide backbone and not side-chains, the anti-parallel β-sheet type binding pattern allows SV3CP to strongly coordinate substrate and is typical of other chymotrypsin-like proteases. Further examination of these and other β-sheet like hydrogen bonding between MAPI main-chain and SV3CP main-chain groups is offered later in this chapter.

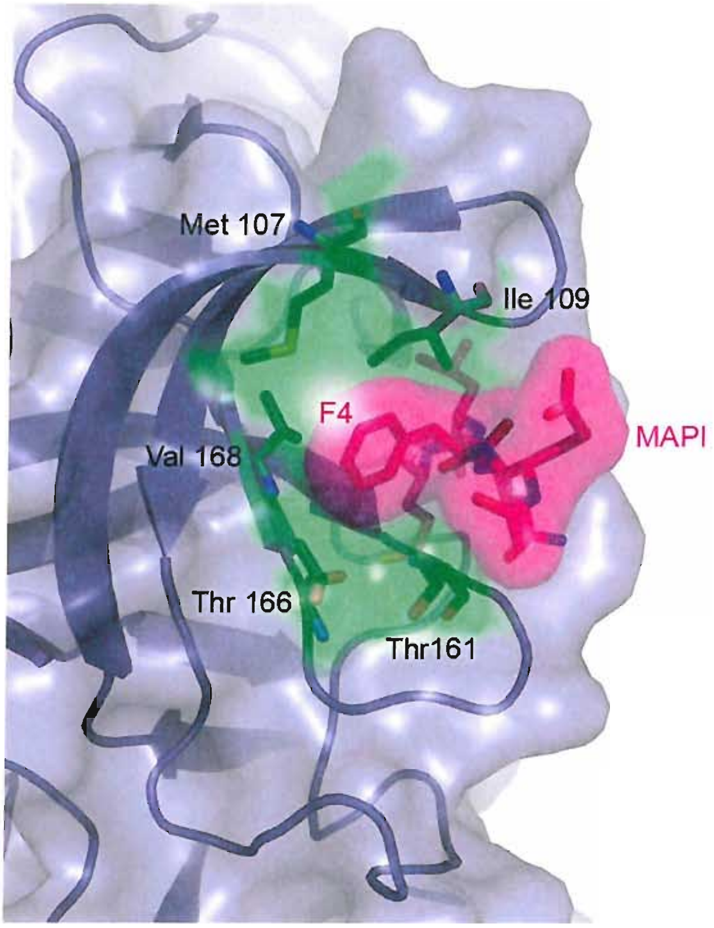


Figure 5.13: The hydrophobic S4 binding pocket. *Laying on the surface of SV3CP, the S4 binding pocket is comprised of the residues: Met 107, Ile 109, Thr 161, Thr 166 and Val 168.*

5.2.3 An active site clamp

Describing the Chiba-virus 3C protease structure, Nakamura *et al* (2005) (170) suggests the side-chains of Gln 110 and Lys 162 act to clamp substrate at the P4 position by forming a bridge under which substrate lies. This hypothesis provides an attractive if possibly flawed model. In support of this 'clamp' model, the SV3CP-MAPI structure shows that neither side-chain is involved in direct interactions with other SV3CP residues thereby allowing the side-chains to flex and the clamp to open and close. It follows that if substrate is *not* bound at the active site cleft the clamp would most probably be in the open position i.e. the Gln 110 and Lys 162 side-chains would not be forming a bridge. The opposite would be true if substrate *was* bound i.e. the clamp would be in the closed position with the side-chains of Gln 110 and Lys 162 forming a bridge over the substrate to brace it within the active site cleft. However, in all three of the *Noro*virus 3C protease structures (SV, Norwalk-virus and Chiba-virus) the clamp is actually seen only in the *closed* position. This is regardless of whether substrate (or substrate analogue) *is* bound at the active site cleft as with the SV3CP-MAPI structure (see *Figure 5.14*), or is *not* bound as with the Norwalk and Chiba-virus structures. It seems clear therefore that for the clamp to adopt the closed position it is not essential for substrate to actually be bound at the active site cleft.

Casting further doubt over the 'clamp' model as presented by Nakamura *et al* (2005), the conformations of Gln 110 and Lys 162 in molecule 1 of the SV3CP-MAPI structure differ from those seen in molecule 2 of the dimer. Whilst both the Gln 110 and Lys 162 side-chains are clamped over F4 of MAPI in molecule 1, they are skewed away from each other and therefore not creating the 'bridge' as seen in molecule 2 of the SV3CP-MAPI dimer. Evidence suggests therefore that the clamp may operate with the side-chains of Gln 110 and Lys 162 in the 'skewed' position as well as the 'bridge' position.

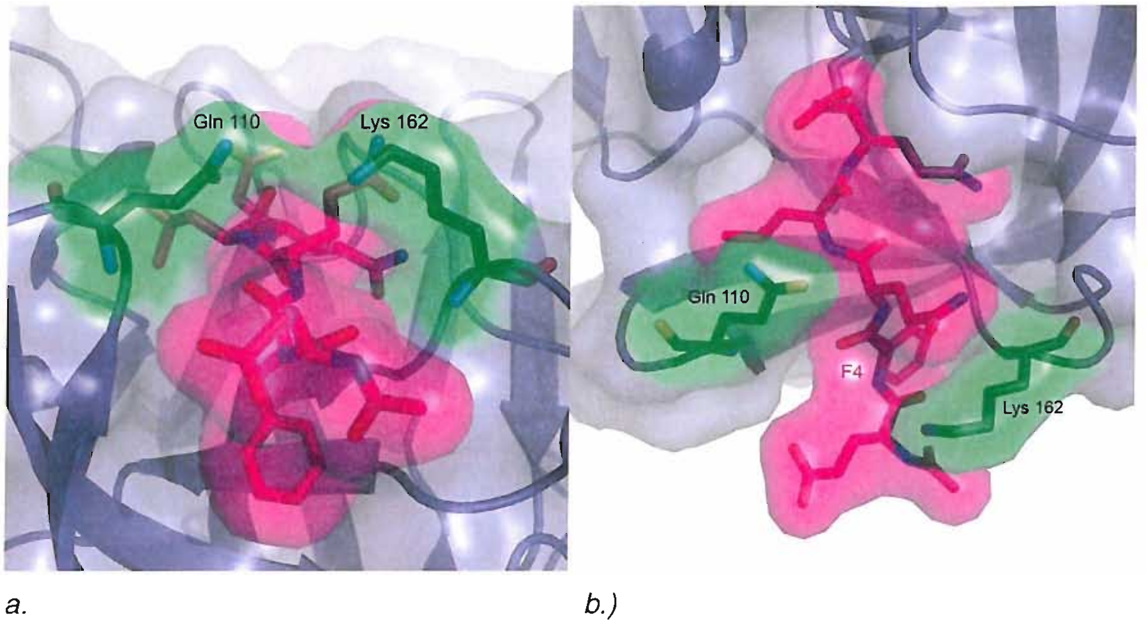


Figure 5.14: The Gln 110/Lys 162 'clamp'. *a.) Shown in the closed 'bridge' position in molecule 2 of the SV3CP dimer, MAPI (shown in pink) is clamped at the F4 residue (equivalent to the P4 residue of the natural substrate) by the side-chains of Gln 110 and Lys 162. b.) Rotated 90° from the equivalent orientation in molecule 2, here the clamp in molecule 1 is shown in the closed 'skewed' conformation. MAPI is still clamped at F4 when the clamp adopts the 'skewed' position.*

5.2.4 Distortion of MAPI Q3 side-chain when clamp is in closed 'bridge' position

Comparison of the conformation of MAPI with the clamp in the closed 'bridging' position, to its conformation when the clamp adopts the closed 'skewed' position reveals a distortion in the MAPI Q3 side-chain. With the clamp in the closed 'skewed' position the conformation of the Q3 side-chain relative to other MAPI side-chains is akin to a peptide chain in an extended conformation. With the clamp in the closed 'bridging' position, the MAPI Q3 side-chain is pushed towards the side-chain of MAPI Q1, so close in fact as to allow hydrogen bonding between the Oδ of MAPI Q3 and the Nδ of MAPI Q1 (see *Figure 5.15*).

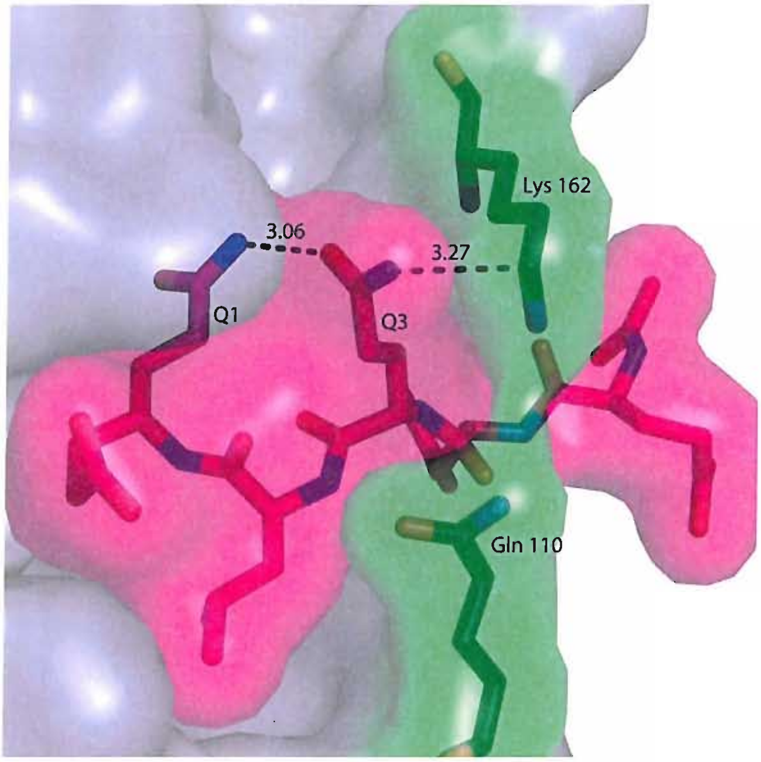


Figure 5.15: Distortion of the MAPI Q3 side-chain when the Gln 110 – Lys 162 clamp adopts the 'bridge' position. *The distorted MAPI Q3 side-chain is brought into sufficient proximity of the MAPI Q1 side as to allow hydrogen bonding between the Oδ of MAPI Q3 and the Nδ of MAPI Q1. The distance between the Lys 162 Cε and the Nδ of MAPI Q3 is also indicated.*

Irrespective of whether the clamp is in the closed 'bridged' position or closed 'skewed' position the MAPI Cα backbone adopts the same conformation. Therefore the distortion of the MAPI Q3 side-chain is due to *local* forces acting upon the Q3 side-chain alone, and is not the result of a more *global* alteration in MAPI conformation. With the clamp in the closed 'skewed' position the MAPI Q3 side-chain shares no interactions with SV3CP (see *Section 5.2.5*). In contrast, with the clamp in the closed 'bridged' position the Nδ of Q3 comes into the *Van der Waal* sphere of the Cε of Lys 162 and as such rotation around the Q3 χ_1 bond is forced resulting in the Q3 side-chain moving from the gauche negative conformation as seen in molecule 1 of the SV3CP dimer to the gauche positive conformation as seen in molecule 2 (see *Figure 5.16*). This gauche positive side-chain conformation is then stabilised by the hydrogen bonding between the Oδ of MAPI Q3 and the Nδ of MAPI Q1 as previously discussed.

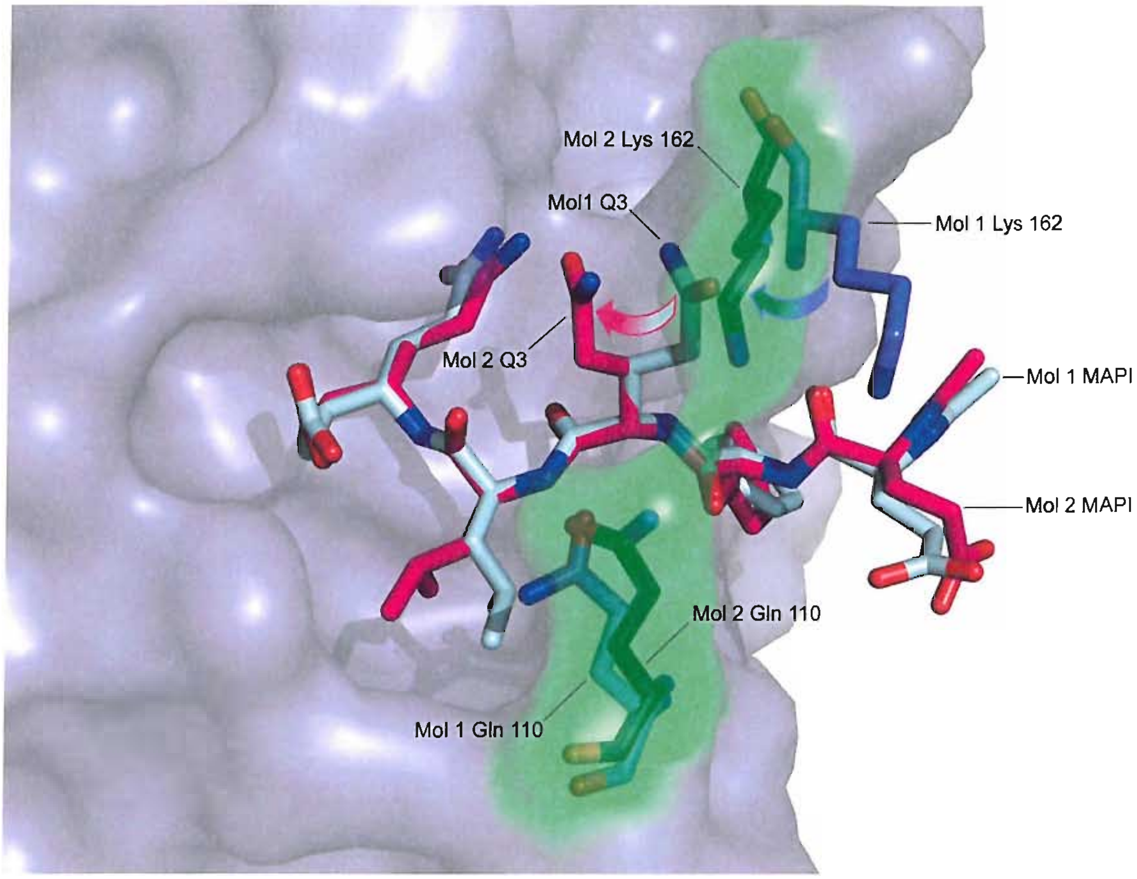


Figure 5.16: The distortion of the MAPI Q3 side-chain by the Gln 110/Lys 162 clamp. *For molecule 1 of the SV3CP dimer, the Gln 110 and Lys 162 side-chains are shown in blue, and MAPI is shown in light blue. For molecule 2 of the SV3CP dimer, the Gln 110, Lys 162 and the corresponding rendered surface is shown in green, with MAPI is shown in pink. Movement of the Lys 162 side-chain from the closed 'skewed' position to the closed 'bridged' position forces rotation around the χ_1 bond of the MAPI Q3 side-chain. The resultant conformation is stabilised by hydrogen bonding between Oδ of MAPI Q3 and the Nδ of MAPI Q1.*

Though interesting by themselves, these observations provide no further explanation as to why the Gln 110/Lys 162 clamp can exist in the two closed conformations. A distortion of the MAPI Cα backbone when the clamp adopts the closed 'bridged' position may have indicated that the natural substrate undergoes an 'induced fit' to better orientate the scissile bond in relation to the SV3CP catalytic triad. However, as previously mentioned the MAPI Cα backbone remains in the same position irrespective of whether the clamp is in the closed 'bridged' or 'skewed' positions so a substrate 'induced fit' clearly does not occur as a result of clamp position. An alternative hypothesis may be that an increased number of SV3CP to substrate interactions would be expected in one of the two closed clamp positions. However, aside from the Q3 to Q1 side-chain hydrogen bonding seen in MAPI with the clamp in the closed 'bridged' position, neither of the clamp positions seems to better co-ordinate substrate through an increased number of SV3CP active site cleft to substrate interactions. In summary, the structural information presently available does not resolve the precise mechanism of the SV3CP active site clamp, even if it does strongly suggest that a substrate clamping mechanism exists.

5.2.5 A large or ill defined S3 binding pocket

Residue identity at substrate position P3 seems unimportant for substrate binding. The SV3CP-MAPI structure clearly shows the side-chain group of MAPI Q3 is not positioned within a defined binding pocket. Further, comparison of residues at position P3 of the five cleavage sites of the natural substrate: His, Gln, Thr, Ser and Thr, reveals characteristically diverse residues. Whilst residue identity may be of little importance at substrate position P3 the anti-parallel β-sheet like hydrogen bonding pattern between the MAPI main-chain carbonyl O and amine N with the SV3CP main-chain Ala 160 amine N and carbonyl O groups respectively (see *Figure 5.17*), is likely to be highly important for binding and orientation of substrate in relation to the active site triad.

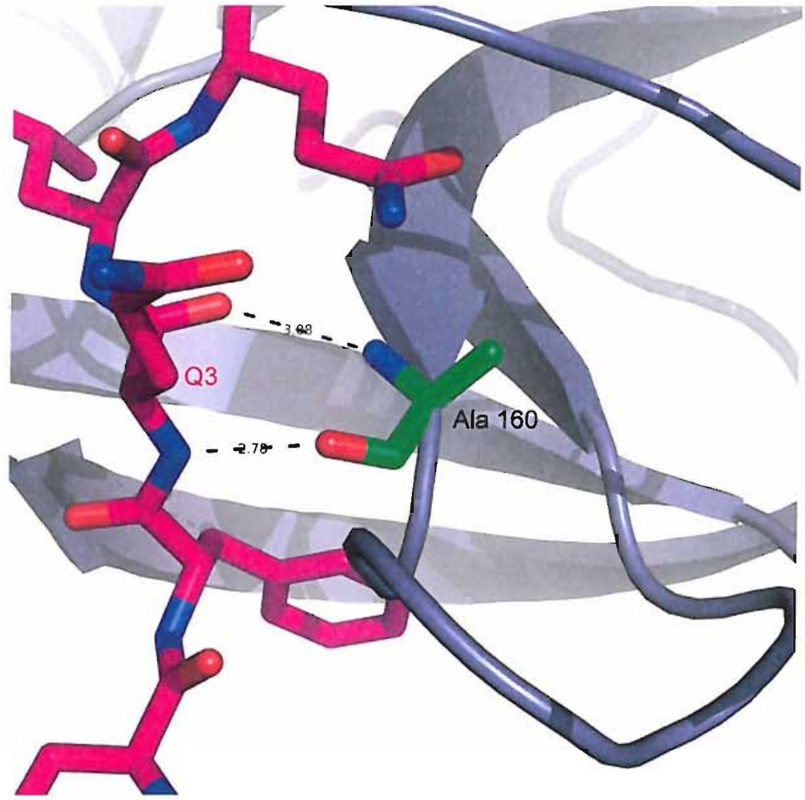


Figure 5.17: The anti-parallel β-sheet like co-ordination of MAPI Q3 with SV3CP*.* *The main-chain carbonyl O and amine N of Q3 of MAPI hydrogen bond in a β-sheet like pattern with the main-chain amine and carbonyl groups of Ala 160.*

5.2.6 The S2 binding pocket

The buried hydrophobic and relatively large S2 binding pocket is occupied by the MAPI L2 side-chain (see *Figure 5.18*). The S2 binding pocket is comprised of the hydrophobic residues: His 30 (of the catalytic triad), Ile 109, Arg 112 and Val 114. The hydrophobic nature of this pocket is further underlined by the absence of any solvent molecules and a complete lack of any electrostatic interactions between the residues of the S2 pocket and the non-polar side-chain of L2; the S2 pocket simply provides a hydrophobic environment in which a hydrophobic substrate side-chain may be accommodated. Comparison of the five points of cleavage in the SV ORF1 polyprotein reveals the residues at the P2 position always possess a hydrophobic side-chain, the residues being: Leu, Leu, Met, Phe and Leu. It seems highly likely that for SV3CP to co-ordinate substrate correctly the P2 residue must be hydrophobic in nature.

Owing to its large size the S2 pocket can accommodate the medium sized hydrophobic residues of Leu and Met, or the relatively large bulky hydrophobic residue Phe (Leu, Met and Phe are seen in the P2 position at the multiple cleavage points of the SV ORF1 polyprotein product). This is demonstrated by comparing the side-chain conformations of L2 between molecule 1 and 2 of the SV3CP-MAPI dimer. In molecule 1 the L2 side-chain is in the gauche positive orientation; in molecule 2 of the dimer, rotation around the χ_1 bond has allowed the L2 side-chain to adopt the trans orientation i.e. the S2 pocket is sufficiently large to permit such free rotation around the χ_1 bond of the L2 side-chain.

Separately from the S2 pocket, the main-chain amine N of L2 of MAPI interacts via hydrogen bonding to the side-chain O of Gln 110. As previously discussed, despite being of importance for substrate coordination, interactions between MAPI main-chain groups and SV3CP (side-chain or main-chain) reveal little information regarding substrate specificity.

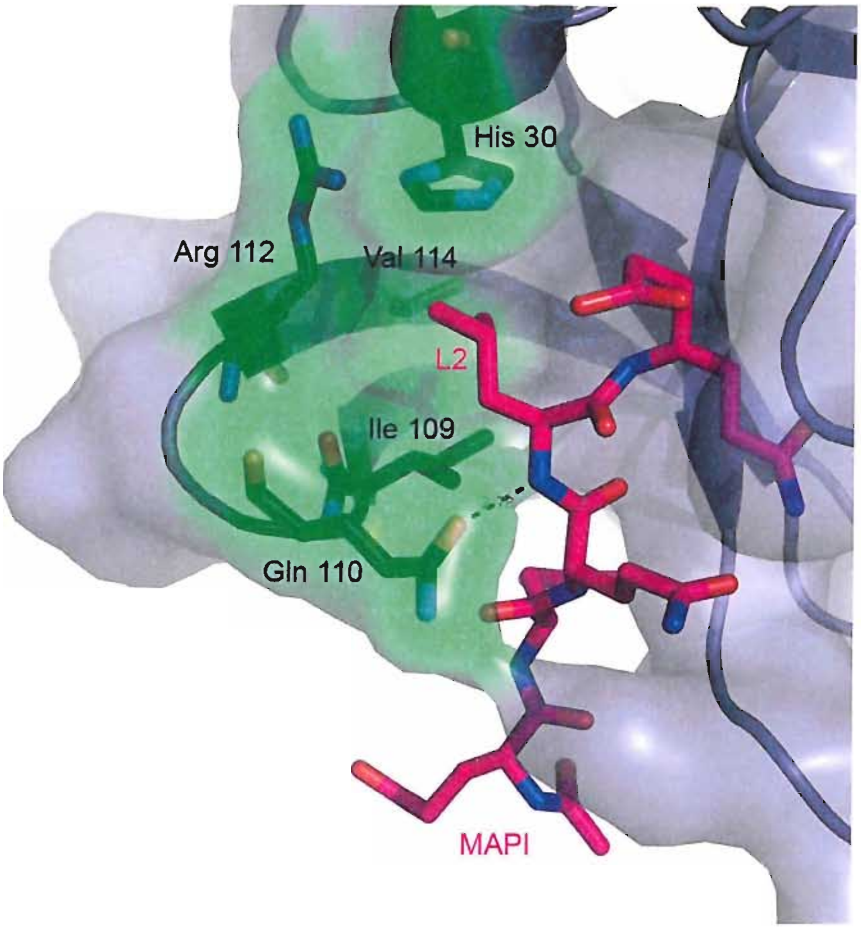


Figure 5.18: The S2 binding pocket of SV3CP. *The buried, hydrophobic S2 binding pocket is comprised of the residues: His 30, Ile 109, Gln 110, Arg 112 and Val 114. The substrate residue at the P2 position must possess a hydrophobic side-chain to be accommodated within the S2 binding pocket.*

Zietler *et al* (2006) observed a 4 Å positional deviation in the residues Ile 109 to Val 114 between their two Norwalk-virus 3C protease structures (a selenomethionine derivative structure and a separate native structure). The SV3CP-MAPI structure shows these residues form the S2 binding pocket and as such, a large degree of flexibility in this region would hold strong implications for substrate binding. Zietler *et al* (2006) attributes the flexibility seen in the Norwalk virus structures to an artefact of crystallisation: that crystal contacts have forced a distortion in this region. However a more likely explanation seems that since neither Norwalk virus structures have substrate (or substrate analogue) bound at the active site cleft, the stabilising effects of protease-substrate interactions upon the residues Ile 109 to Val 114 are absent and therefore this region is able to flex. In contrast the SV3CP-MAPI structure shows no such flexibility in this region, presumably since the stabilising interactions between protease and substrate absent in the Norwalk virus structures are present in the SV3CP-MAPI structure.

5.2.7 The S1 binding pocket and MAPI Cys139 covalent linkage

The S1 and S1' (S1' discussed in Section 5.2.9) binding pockets arguably play the most important role in substrate sequence recognition and substrate orientation. Comparison of the five di-peptide sequences that occupy the S1 and S1' sites: QG, QG, EG, EA and EG, reveals SV3CP will accept only a Gln or Glu at the S1 site, and only a Gly or Ala at the S1' site. Inspection of the SV3CP-MAPI structure shows the S1 site is relatively large in size and able to accommodate the Q1 side-chain of MAPI in a fully extended conformation. It is clear from the biochemical evidence alone that the S1 site differs quite dramatically from the S1' binding site in its ability to accommodate a relatively large substrate side-chain; the S1' site seems only able to accommodate the small side-chains of glycine and alanine.

The S1 binding site seems less easily definable as the S2 and S4 sites, though the Q1 side-chain can clearly be seen to lay parallel to, and be sandwiched between the loop: lpeII, and the C-terminal region of the β-strand: βeII of SV3CP. However, the cleft made by lpcII and βeII (*figure 5.19*) simply provides sufficient space into which the relatively long, but non-bulky side-chains of the Gln or Glu substrate P1 side-chains may be accommodated without actually sharing electrostatic interactions, directly or via water molecules with residues of the S1 binding site. There are however two very important exceptions to this observation. The residues Thr 134 of lpeII and His 157 of βeII are located at the end of the S1 binding pocket distal to the MAPI Cα backbone and as such are positioned to hydrogen bond with the head groups of the substrate side-chains: Gln or Glu (the only residues found in the natural substrate at the P1 position). Thr 134 and His 157 are therefore of great importance in co-ordination of substrate and in defining substrate specificity at the SV3CP S1 site (see *Figure 5.19*). The essential role played by His 157 to substrate binding at the S1 site has been demonstrated by mutational studies (171) that showed a His157Ala mutant displays a reduced rate of substrate turnover presumably due to a reduction in the efficiency of substrate binding. It also appears that Tyr 143 may play an important role in substrate interactions at the S1 site, though not directly. The SV3CP-MAPI structure suggests that for correct orientation of the His 157 side-chain, the Cζ OH group of Tyr 143 must hydrogen bond to the Nδ1 of His 157. This interaction stabilises the His 157 side-chain presumably so His 157 is better orientated to hydrogen bond to the substrate side-chain occupying the S1 site.

Inspection of the SV3CP-MAPI electron density maps reveals a lack of density for the ethyl ester extension of MAPI. It is clear that the ethyl ester extension has been hydrolysed to yield a carboxylic acid group. It is with this carboxylic acid group that the SV3CP residues: His 30 and Gly 137 interact. The Nε2 of His 30 hydrogen bonds to the O2 of the MAPI carboxylic acid. This interaction is likely to be restricted to the SV3CP-MAPI structure only due to the distorted position of His 30 (see *Section 5.2.8*). The Gly 137 main chain amine hydrogen bonds to the O1 of the MAPI carboxylic acid as it would when behaving as a member of the oxyanion hole. This is because the SV3CP-MAPI structure represents a single fixed state, where MAPI is irreversibly bound to Cys 139 of the SV3CP active site and not cleaved as the natural substrate would be. The transient hydrogen bonding that would normally exist between Gly 137 behaving as a member of the oxyanion hole and substrate during proteolysis can therefore be observed in the SV3CP-MAPI structure.

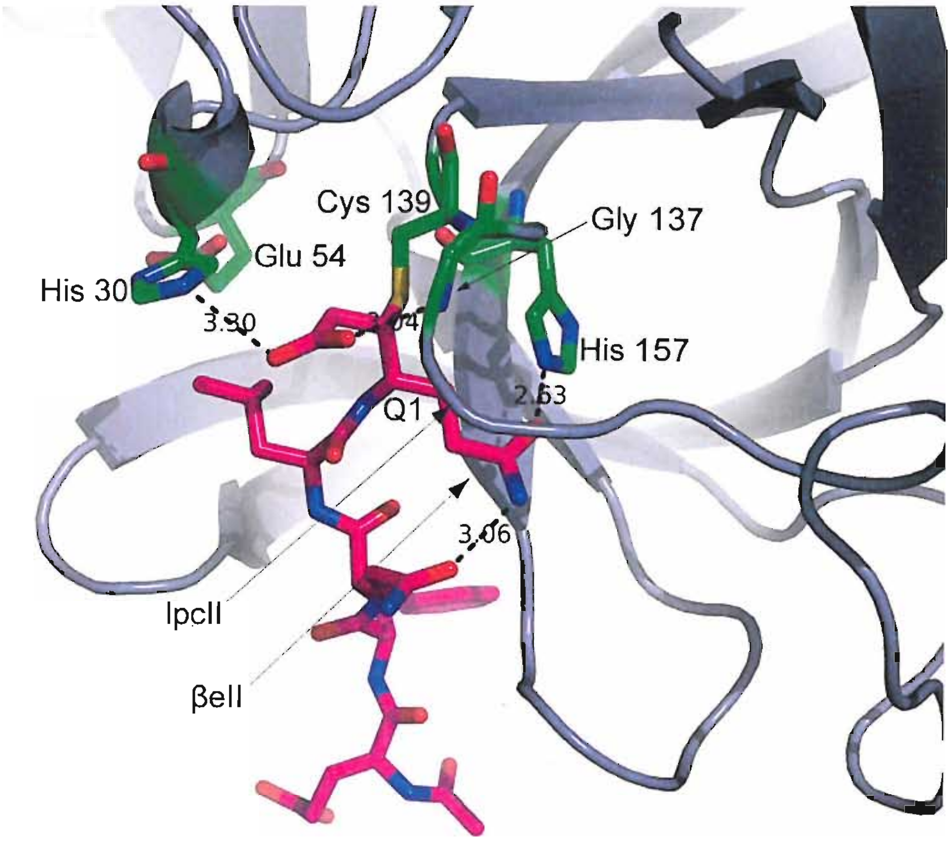


Figure 5.19: The S1' binding site of SV3CP. *The members of the SV3CP catalytic triad: His 30, Glu 54 and Cys 139 are highlighted in green, as are the residues involved in co-ordination of MAPI (and presumably natural substrate) at the S1 binding site: Gly 137 and His 157. Gly 137 is also one member of the SV3CP active site oxyanion hole (see Section 5.3). The loop: lpcII, and the beta strand: βeII form a cleft that can accommodate the side chain of the P1 residue.*

Inspection of electron density between the Sγ of Cys 139 and MAPI suggests these atoms are covalently bonded (see *Figure 5.20*). This observation is in agreement with the mass spectroscopy data and kinetic assay work (see *Section 4.4.9.3*) that both indicate SV3CP is irreversibly modified with MAPI. Indeed, MAPI was designed to do just that: to covalently bond to SV3CP to cause the irreversible inhibition of protease activity. This observation provides structural, and in combination with the mass spectroscopy and kinetic assay work, conclusive evidence of the efficacy of MAPI as an irreversible inhibitor of SV3CP.

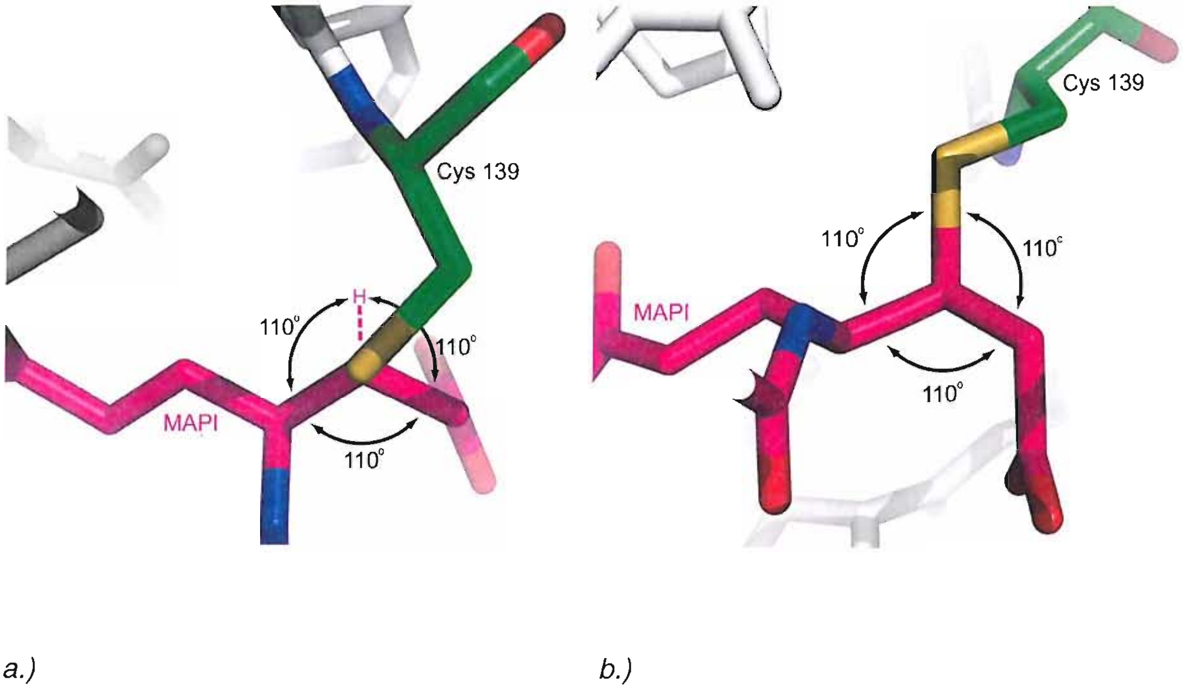


Figure 5.20: The covalent linkage between Sγ of the catalytic Cys 139 of SV3CP and MAPI. *There is a tetrahedral distribution of bonds around the C of MAPI to which the Sγ of Cys 139 is linked by a covalent bond. a.) Looking along the Sγ to MAPI C covalent bond. The H of covalently bound MAPI C is shown in pink. b.) Rotated 90° to a.).*

5.2.8 Distortion of the active site triad upon SV3CP modification with MAPI

In both molecules of the SV3CP-MAPI dimer, the side-chain of His 30 is separated by too great a distance from the side-chains of Glu 54 and Cys 139 to allow the necessary hydrogen bonding required to form the catalytic triad (see Section 5.3). With MAPI covalently bound to the Sγ of Cys 139 the Nδ1 of His 30 can no longer hydrogen bond to the former Cys 139 sulfhydryl group allowing the His 30 side-chain to adopt a conformation pointing away from Cys 139 and Glu 54. Presumeably the His 30 side-chain can flip away from Cys 139 since the hydrogen bond that would normally exist (in the absence of MAPI) is lost and the His 30 side-chain is no longer tethered to Cys 139. With natural substrate bound the His 30 side-chain would adopt a conformation that is within hydorgen bonding distance of Cys 139 and Glu 54. Observing stereochemical restraints of His side-chain orientation, such a conformation is proposed in *Figure 5.21* and places the His 30 Nε2 and Nδ1 2.74 Å and 3.04 Å from the Glu 54 Oδ and Cys 139 Sγ respectively, i.e. within hydrogen bonding distances. This simple modelling of the His 30 side-chain agrees with the biochemical evidence (171) that it is indeed His 30 and not His 157 (that is also within the proximity of the catalytic Cys 139) that participates in the active site triad.

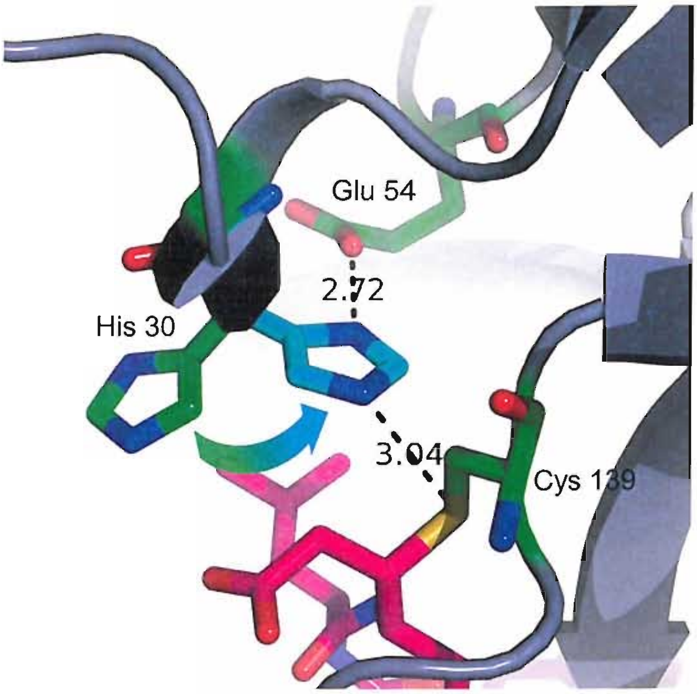


Figure 5.21: Proposed alternate conformation (coloured cyan) of His 30 side-chain as a member of the catalytic triad during proteolysis of the natural substrate. *The SV3CP-MAPI structure shows the His 30 side-chain in the conformation coloured green; in this orientation both the His 30 Nδ1 and Nε1 are positioned too far from Glu 54 and Cys 139 side-chains to complete the active site triad.*

5.2.9 A prediction of the substrate binding S' sites of SV3CP

As discussed, there exists a possibility that the flexible loop region comprised of the residues Thr 123 to Asp 131 may contribute to substrate binding at SV3CP S' sites. The hypothesis being that the flexibility in this region observed in the SV3CP-MAPI structure is attributable to the absence of substrate (or analogue i.e. MAPI) occupying the S' sites and the resultant lack of stabilising substrate to SV3CP interactions. However, in light of the presently available structural information that displays this region in a disordered state any prediction of S' sites owing to this loop cannot be made. Further, visual inspection of a surface rendered model of the SV3CP-MAPI structure suggests substrate P' residues may be co-ordinated elsewhere providing further evidence that this disordered region, as previously postulated, is involved in SV3CP dimer interactions and not substrate binding.

Inspection of *Figure 5.22* displaying a surface rendered depiction of the SV3CP-MAPI model, reveals what appears to be a rather deep surface cleft that follows on from the S site binding cleft occupied by MAPI. It seems quite possible that this unoccupied cleft may co-ordinate substrate P' residues. If so, the S1' binding site may consist, in whole or in part, of the residues Ser 14, Val 31 Gly 137 and the catalytic residue Cys 139. Biochemical evidence suggests the P1' substrate residue of SV3CP always possesses a small side-chain group, therefore requiring only a small S1' pocket in which to be accommodated. The relative close proximity of Ser 14, Val 31, Gly 137 and Cys 139 would provide such a small sized S1' binding pocket.

Subsequent S' sites are less easy to define though it is possible to suggest two putative routes along which substrate P' residues may lie. The most obvious route follows the deep surface cleft mentioned in the previous paragraph, this is named 'route I' in *Figure 5.22.* If substrate were to lie along this route, the residues of SV3CP that may interact with substrate are: Lys 11, Ser 14, Trp 16. Arg 89, Asp 90, Ser 91, Ile 135, Gly 137 and Asp 138. Alternatively substrate may lie along the route named 'route II'. Though whether substrate lies along route II is dependent upon the true nature of Lys 11 and Asp 90. From visual inspection of the SV3CP-MAPI model it seems possible that these residues may operate as a clamp in a similar manner to the Gln 110/Lys 162 clamp. If so, the putative Lys 11/Asp 90 clamp may coordinate substrate at the substrate P3' or

P4' residues. However, this hypothesis as with all hypotheses regarding substrate binding at the S' sites of SV3CP remains pure speculation in the absence of the structural proof provided by a structure of SV3CP with substrate (or substrate analogue) occupying the S' sites. Such a structure has yet to be determined for a 3C protease of any *Noro*viral strain.

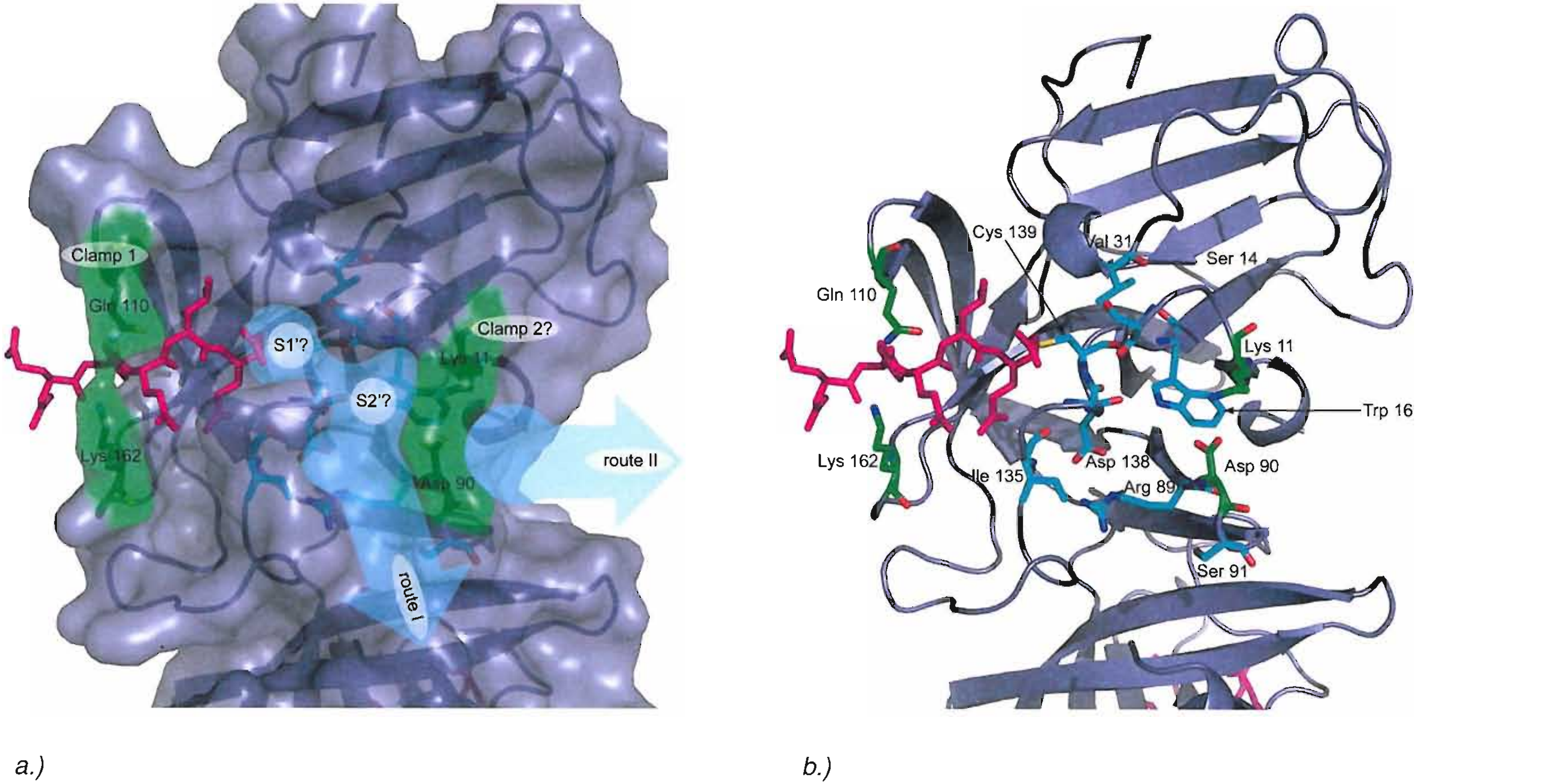


Figure 5.22: The predicted S' substrate binding sites of SV3CP: *a.) Highlighted in cyan are the predicted locations of the S' substrate binding sites of SV3CP. The S1' site is predicted to be comprised of the residues: Ser 14, Val 31 Gly 137 and Cys 139. Substrate P' residues may lie along one of two routes labelled: 'route I' and 'route II'. The SV3CP residues predicted to form the remaining S' sites of 'route I' are highlighted in cyan. If substrate P' residues were to lie along 'route II', the substrate may be clamped at the P3' or P4' by the putative clamp formed of Lys 11 and Asp 90. b.) For clarity the rendered surface has been removed and residues highlighted in a.) are labelled. MAPI is shown in pink throughout.*

5.2.10 A general β-sheet like hydrogen bonding pattern to co-ordinate substrate

Typical of chymotrypsin proteases, SV3CP co-ordinates MAPI via a hydrogen bonding network in the mode of an anti-parallel β-sheet (see *Figure 5.23*). Although such interactions involving the main-chain atoms of the substrate (or substrate analogue i.e. MAPI) are not responsible for defining protease specificity they are of great importance in correctly co-ordinating substrate in relation to the protease catalytic residues. Some of these interactions have already been discussed earlier in this chapter but will be re-examined as a whole to gain a fuller appreciation of this protease to substrate β-sheet like binding pattern.

The key interactions are between Ala 160 and MAPI Q3, and, Ala 158 and MAPI Q1. The MAPI Q3 carbonyl O and amine N hydrogen bond to the amine N and carbonyl O of Ala 160 respectively, in typical fashion of a hydrogen bonding network of an anti-parallel β-sheet. A further *main-chain to main-chain* hydrogen bond exists between the carbonyl O of Ala 158 and the amine N of MAPI Q1. These three *main-chain to main chain* interactions must be considered the most important of the SV3CP to MAPI main-chain interactions since they are involve main-chain atoms, not side-chain atoms of SV3CP. The remaining hydrogen bonding between MAPI and SV3CP, that may loosely be considered to follow a β-sheet bonding pattern, involves MAPI main-chain atoms hydrogen bonding to side-chain atoms of SV3CP either directly or via intermediate water molecules. These interactions include: the amine N of MAPI Q3 to the Oδ of Gln 110, the carbonyl O of MAPI Q4 to the amine N of Gln 110 via a water molecule, and, the amine N of MAPI Q4 to the carbonyl O of Arg 108 also via a water molecule.

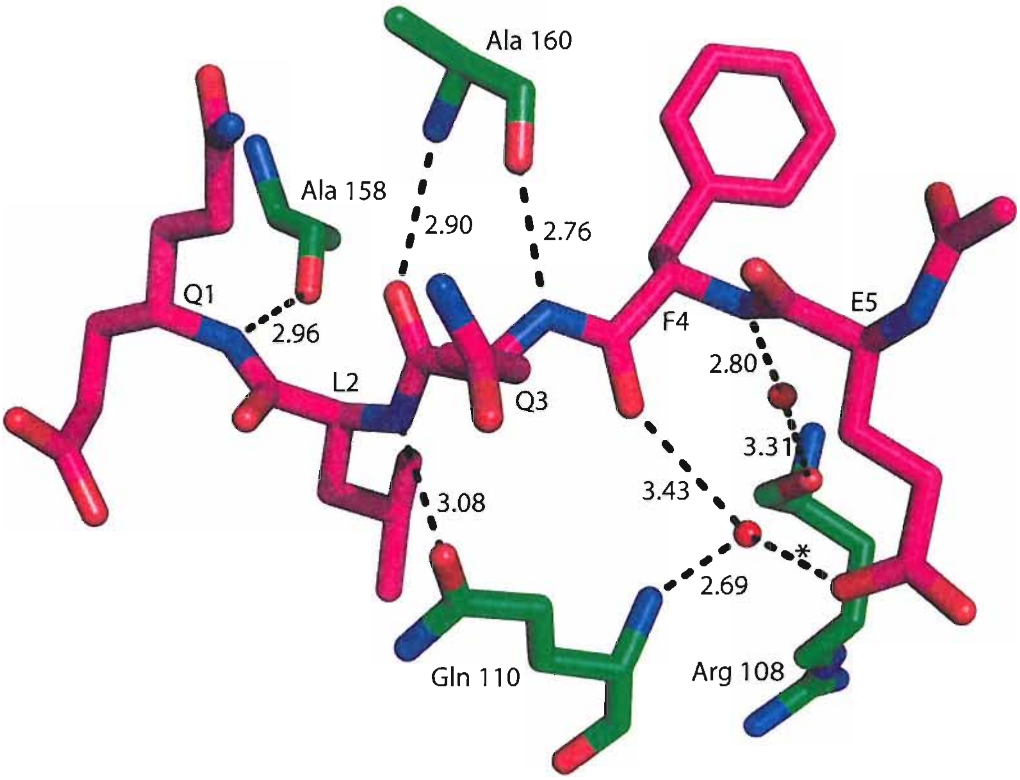


Figure 5.23: The β-sheet like hydrogen bonding network between active site cleft residues of SV3CP and MAPI. *SV3CP residues are shown in green. MAPI is shown in pink. Bond distances are in Å. *though not part of the β-sheet like hydrogen bonding pattern this hydrogen bond is included in the diagram to show how the water molecule to which it is hydrogen bonded is coordinated to two further groups: the amine N of Gln 110 of SV3CP and the Oδ1 of E5 of MAPI.*

5.3 Proposed mechanism of substrate cleavage

Based upon the SV3CP structure and the widely accepted mechanism of catalysis by cysteine proteases, a specific mechanism may be proposed through which substrate peptide bond hydrolysis by SV3CP proceeds. A schematic representation of the proposed mechanism is shown in *Figure 5.25: parts 1 and 2.*

As previously identified following inspection of the SV3CP-MAPI structure, the members of the active site catalytic triad are: Cys 139, His 30 and Glu 54 and, based upon this observation, the following mechanism of catalysis may be proposed. Glu 54 immobilises

His 30 via hydrogen bonding between the Glu 54 Oδ1 and the His 30 Nδ1 so that His 30 acting as a Brønsted-Lowry base may de-protonate Cys 139 (*Figure 5.25: part1: b*). With the correct orientation of the substrate (the SV ORF1 poly-protein) Cα backbone within the active site cleft, the substrate scissile bond is brought into sufficient proximity of the active site Cys 139 to facilitate nucleophilic attack by the Cys 139 on the carbonyl C of the substrate scissile bond to yield an unstable tetrahedral intermediate where substrate is covalently bound to SV3CP (*Figure 5.25: part1: b. and c.*). The de-protonation of Cys 139 by His 30 and nucleophilic attack on the carbonyl C are likely to be concurrent events as nucleophilic attack on the substrate carbonyl C alone would leave the Cys 139 side-chain Sγ with a negative charge and quite reactive as seen in the papain family of proteases. During nucleophilic attack on the substrate carbonyl group the shift of negative charge onto the substrate carbonyl O is stabilised by hydrogen bonding of the carbonyl O to the SV3CP Cα backbone amine groups of Cys 139 and Gly 137 that form the *oxyanion hole* (*Figure 5.25: part1 c.*). It is proposed that Asp 138 is *not* a member of the oxyanion hole as reported by Zietler *et al* (2006) though is vital in orientation of the oxyanion hole members though through indirect means. The main-chain amine N of Asp 138 hydrogen bonds to the main-chain carbonyl O of Ile 135 thereby stabilising the tight hairpin loop that correctly orientates the main-chain amine group of Gly 137 towards the main-chain amine of Cys 139 to form the oxyanion hole.

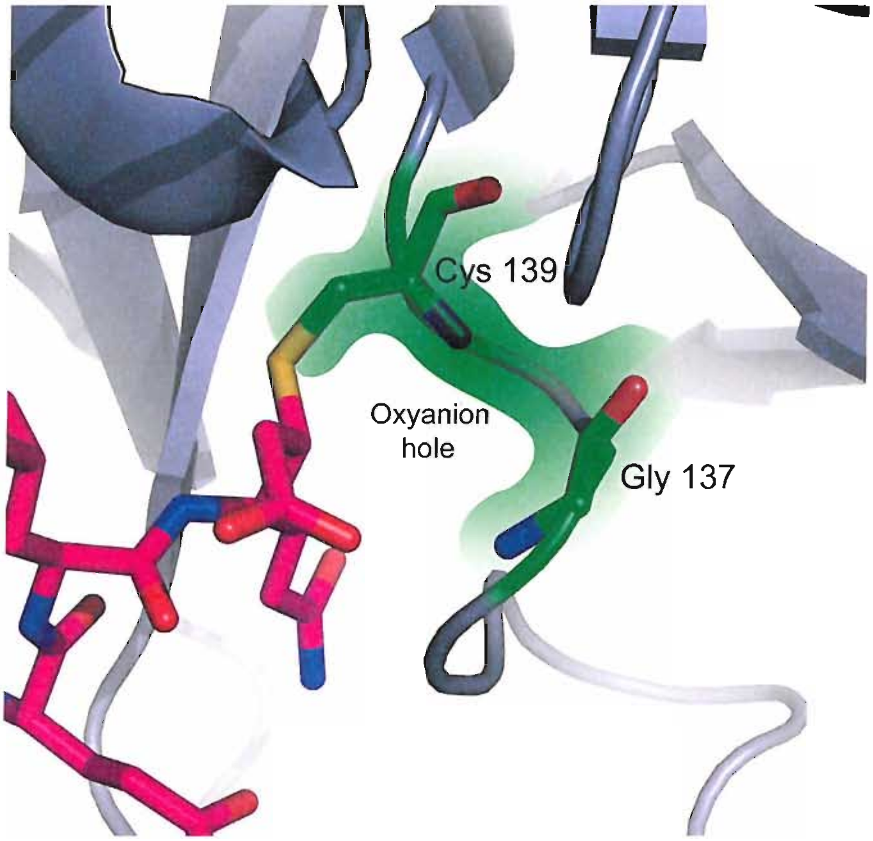


***Figure 5.24: The SV3CP oxyanion hole**. The main-chain amide groups of Cys 139 and Gly 137 comprise the oxyanion hole that during cleavage of natural substrate would stabilise the carbonyl O of the substrate scissile bond by hydrogen bonding during formation of the first and second tetrahedral intermediate states of substrate cleavage. MAPI (shown in pink) possesses a modified C-terminus and so lacks a S1 carbonyl group, however the C of the Michael acceptor extension of MAPI that undergoes nucleophillic attack by the sulfhydryl group of Cys 139 of the SV3CP active site is analogous to the carbonyl C of the natural substrate. Though of course due to the nature of the Michael acceptor extension this C of MAPI does not possess an O and therefore the oxyanion hole is not occupied in the SV3CP-MAPI structure. However despite this rather significant difference, the orientation of this C of MAPI within the active site strongly indicates that the main-chain amide groups of Cys 139 and Gly 137 form the oxyanion hole.*

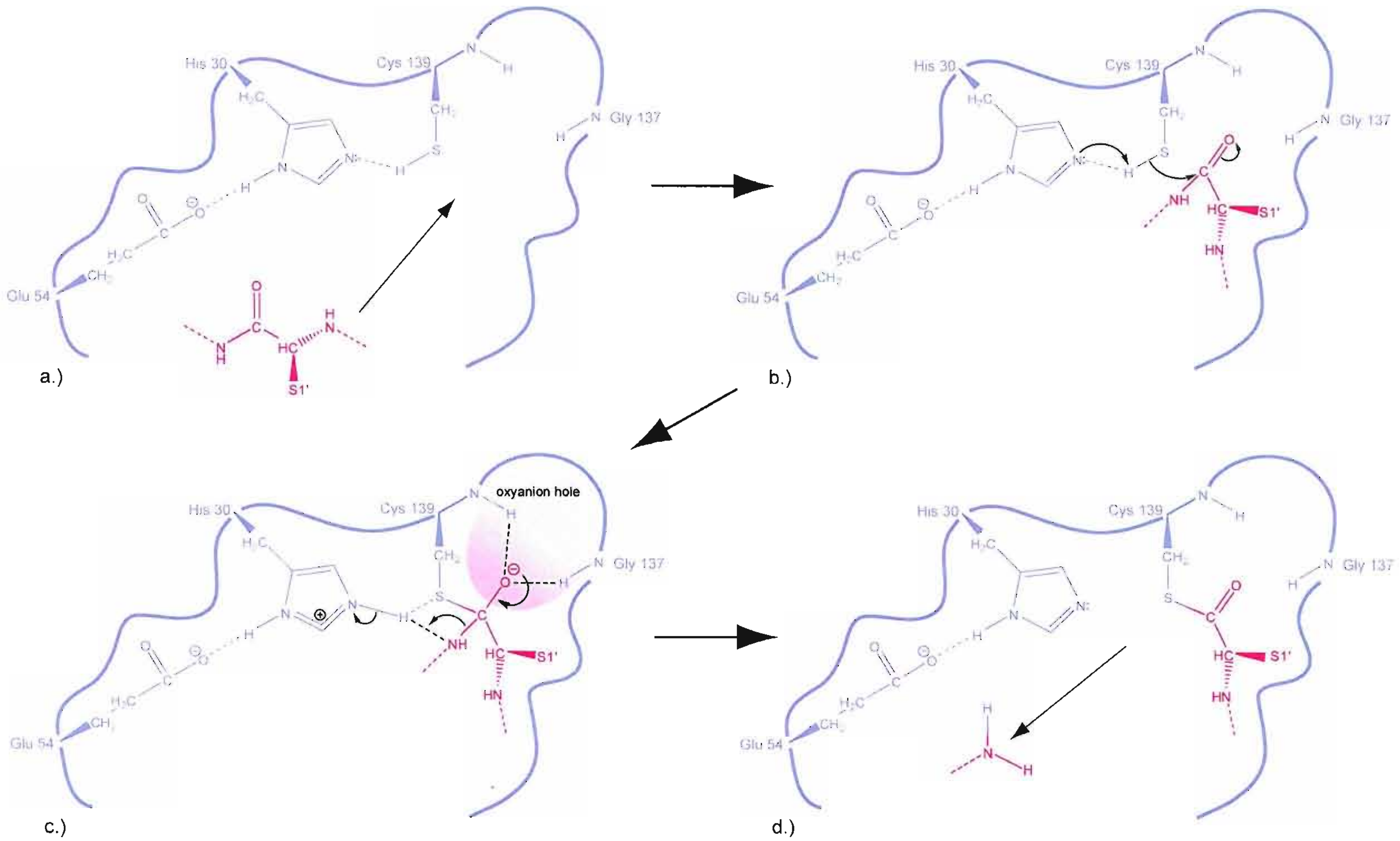


Figure 5.25: Part 1: Proposed mechanism of substrate proteolysis by SV3CP*. Formation of the acyl-enzyme intermediate via a tetrahedral intermediate and elimination of the amine product.*

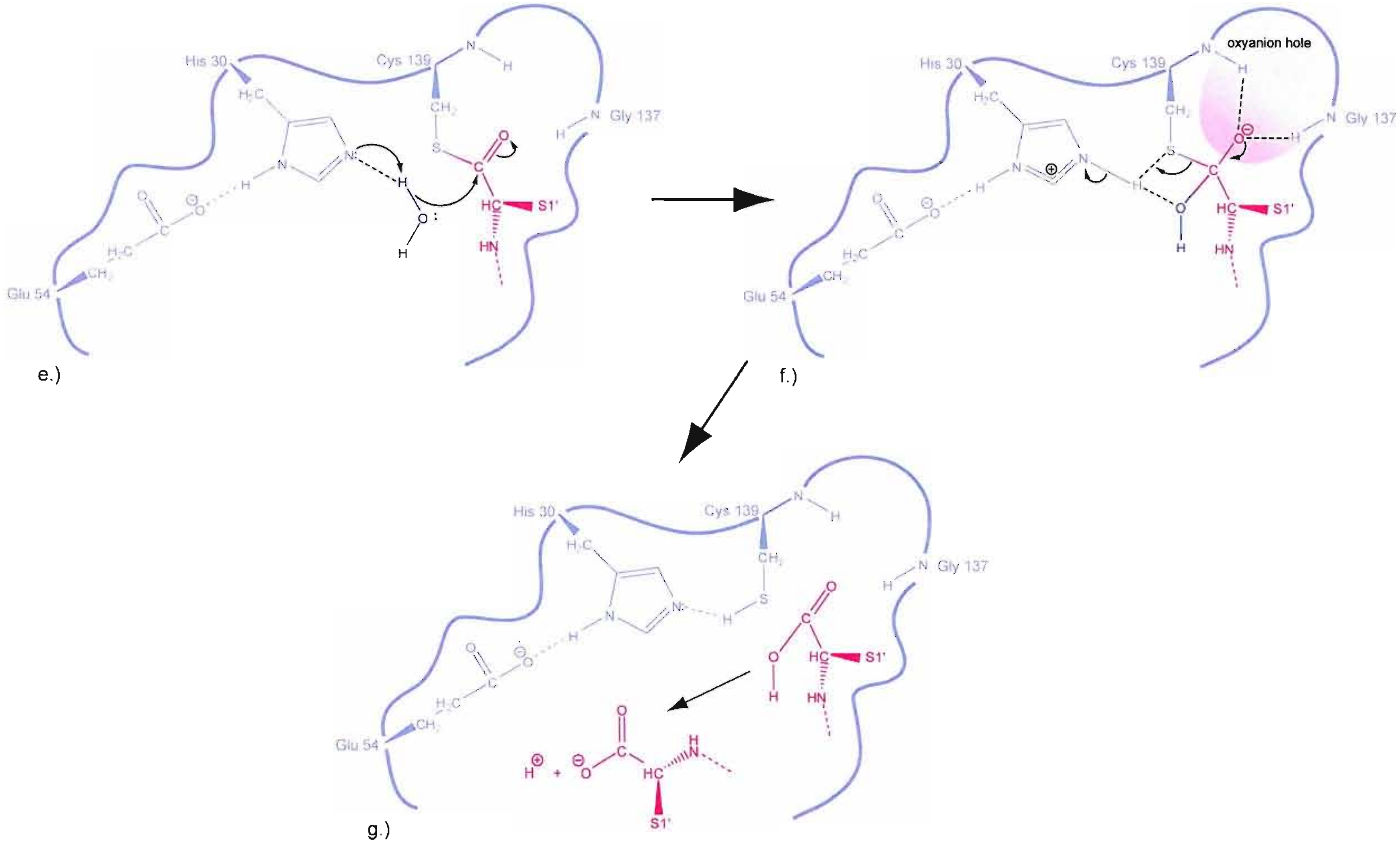


Figure 5.25: Part 2: Proposed mechanism of substrate proteolysis by SV3CP. *Hydrolysis of the acyl-enzyme intermediate via a tetrahedral intermediate and generation of the carboxylic acid product.*

Collapse of the tetrahedral intermediate complex is promoted by His 30, now acting as a Brønsted-Lowry acid, protonating the substrate amide N. Protonation of the amide N induces a shift of charge from the negatively stable carbonyl O resulting in the rapid collapse of the tetrahedral intermediate (*Figure 5.25: part1: c. and d.*) and simultaneous re-formation of the substrate carbonyl C=O. The substrate scissile amide bond is then broken and the amine substrate product (the N-terminal portion of the cleaved substrate) dissociates from the active site cleft i.e. the substrate is cleaved (*Figure 5.25: part1: d.*). Steric hindrance between the amine product and the reformed carbonyl C=O has been suggested to expedite the exit of the amine product. The carbonyl carbon of the former substrate scissile bond remains covalently linked to the Cys 139 side-chain Sγ in a relatively stable state termed the *acyl-enzyme intermediate* (*Figure 5.25: part1 d.*).

Following formation of the *acyl-enzyme* intermediate it becomes a case of regeneration of the SV3CP by bringing about the release of the covalently bound C-terminal portion of the substrate. The *acyl-enzyme* state permits the diffusion of a single water molecule into close proximity of the active site triad (*Figure 5.25: part2 e.*). Acting as a Brønsted-Lowry base His30 is protonated by the incoming water molecule allowing the water molecule to nucleophilically attack the substrate carbonyl carbon resulting in a second tetrahedral intermediate where once again the negatively charged carbonyl O is stabilised through hydrogen bonding with the member amine groups of the oxyanion hole (*Figure 5.25: part2: f.*). Concurrently His 30 behaving now as a Brønsted-Lowry acid protonates the Cys 139 side-chain Sγ causing the collapse of the second tertiary intermediate, the release of the C-terminal portion of the substrate peptide and complete regeneration of the active site triad ready for further substrate cleavage (*Figure 5.25: part2: f. and g.*).

5.4 Closing comments

The original aims of this research project, namely: i.) the development of a substrate based colorimetric assay, ii.) the design and synthesis of a potent and specific inhibitor of SV3CP activity, and, iii.) the structural solution by X-ray crystallography of SV3CP in complex with the inhibitor have been achieved.

Following successful synthesis and purification of a series of chromogenic peptides it was possible to complete the kinetic analysis presented in this thesis that proved sufficient to reveal a maximal rate of substrate turnover. The maximal rate of substrate turnover was achieved when SV3CP was able to coordinate residues in the S1 to S5 positions. This is demonstrated by a maximal rate of substrate turnover with the penta-peptide chromogenic substrate: Ac-EFQLQ-*p*NA.

The results of the kinetic analysis allowed for the design of a highly potent and specific inhibitor of SV3CP that comprised of the same penta-peptide sequence as the chromogenic substrate: EFQLQ, but included a C-terminal modification in the form of a Michael acceptor ethyl-ester extension. The Michael acceptor portion of this modified peptide: MAPI, was designed to undergo nucleophillic attack by the sulfhydryl group of the SV3CP active site cysteine: Cys 139, and so become covalently and therefore irreversibly bound to Cys 139 rendering SV3CP inactive. The penta-peptide portion of MAPI was included so that MAPI would display specificity for Cys 139 only, and not the other 4 cysteine residues of SV3CP. Following successful synthesis and purification of MAPI, assay of SV3CP in the presence of the chromogenic substrate: Ac-EFQLQ-*p*NA and the putative inhibitory compound: MAPI, revealed that SV3CP became rapidly inactivated in the presence of, and therefore presumably through modification by, MAPI as SV3CP was unable to turn over the chromogenic substrate. Mass spectroscopy later proved that each molecule of SV3CP became modified by only a single molecule of MAPI. Modification of SV3CP by MAPI was presumed to be via covalent bond to the active site Cys 139. Empirical proof by *trypsin digest* and *peptide mass finger printing* that it was indeed Cys 139 that was modified with MAPI was not sought owing to the rather robust evidence provided by the combination of the kinetic analysis with chromogenic substrate and *exact-mass* mass spectroscopy. It was assumed that in

MAPI, a highly potent and specific inhibitor of SV3CP had been found. The SV3CP-MAPI structure would later prove these assumptions correct.

A structural solution by X-ray crystallography was then sought for SV3CP in complex with MAPI to allow dissection of the details of substrate binding at the SV3CP active site cleft to allow conclusions to be drawn as to SV3CP substrate specificity and possibly the mechanism of substrate cleavage. It was anticipated that such a structure would be of use in the structure led design of further inhibitory compounds of SV3CP. MAPI was the ideal compound to achieve these aims as not only did it include a peptide portion that would mimic the natural substrate P1 to P5 residues to allow for the SV3CP substrate binding sites and their interactions with substrate to be defined, but due to the Michael acceptor ethyl ester extension, would be irreversibly bound to SV3CP thereby circumventing the difficulties of co-crystallising protein and other ligand that may only transiently associate. MAPI served two important purposes, i.) it inhibited SV3CP activity *in vitro*, and also, ii.) proved essential in obtaining an informative structural solution of SV3CP.

Due to lack of suitable SV3CP homologues at the time of structural solution (2 *Noro*virus 3C protease structures: of the Norwalk and Chiba viruses, have since been published), phases were sought experimentally by MAD techniques. A structure of SV3CP in complex with MAPI was finally obtained to a resolution of 1.75 Å with an *R* factor of 20.46 and an R_{free} of 22.27. The SV3CP-MAPI structure has allowed for the SV3CP dimer interface to be defined, the S1 to S5 substrate binding sites to be fully described and putative S' binding site to be suggested.

5.5 Further work

Some issues relating to the mechanisms of substrate binding and cleavage by SV3CP remain. These issues are primarily as a result of the lack of MAPI residues analogous to substrate P' residues. Of course the very nature of MAPI, i.e. the functional C-terminal ethyl-ester extension, prevents MAPI from including such residues. Correspondingly the specifics of substrate binding at SV3CP S' binding are not evident in the SV3CP-MAPI structure. One solution would be to co-crystallise SV3CP with a peptide comprising of the natural substrate residues spanning from the P5 residue to the P4' or P5' residue (or the predicted limit of substrate coordination at SV3CP S' sites). Though using a non-modified peptide such as this would mean forfeiting the benefits of having ligand irreversibly bound to SV3CP as provided by MAPI and so introduce the inherent difficulties of obtaining crystals of protein in complex with a peptide ligand. A further possible solution would be to synthesise a compound similar to MAPI though to include an N-terminal instead of a C-terminal Michael acceptor extension. That way the peptide portion, likely to be no longer than 5 or 6 residues, of the N-terminally modified MAPI-like compound could comprise of the natural substrate P' residues. By this approach the present SV3CP-MAPI structure in combination with a structural solution of SV3CP in complex with the proposed N-terminally modified MAPI-like compound would provide the detail of substrate binding at both the S and S' SV3CP substrate binding sites.

On a non-crystallographic theme, further work is also required to reduce the largely peptide characteristic of MAPI whilst still retaining its specificity. Bulky, charged peptide sequences are often troublesome to translocate across cell membranes, hence drugs sharing such characteristics often possess a low bioavailability. If a drug designed to inhibit viral replication can not access the replicating virus inside its host cell it is quite obviously of little use therapeutically. The Michael acceptor ethyl-ester group of MAPI has been demonstrated *in vitro* to inhibit SV3CP. However in order to be of use in tackling SV infection the peptide portion, as discussed, must be altered to be more amenable to passing across cell membranes whilst still retaining specificity for the SV3CP active site cleft. It is apparent that the less specific the altered MAPI becomes the more likely it is to inhibit host cell proteases, an obvious negative side effect. The peptide portion could be modified to become less hydrophilic and so pass more easily across the mainly hydrophobic host cell membrane. However, drug solubility must be retained, and taking this proposal to its extreme would result in a hydrophobic compound of low solubility.

The MAPI ethyl-ester functional group has proved to be an effective inhibitor of SV3CP and therefore presumably will cause an arrest of SV replication. Ultimately however, to arrive at a suitable solution to improving the availability of a MAPI based SV3CP inhibitor to the SV host cell may require screening of a multitude of MAPI peptide replacements with replicating virus. Presently SV is proving refractory to *in vitro* replication, though hopefully this situation will soon change.

Finally, investigation by mutation of single, and multiple SV3CP residues may elucidate the mechanism of the Gln 110/Lys 162 clamp, whether the putative Lys 11/Asp 90 clamp does clamp substrate in the S' sites, and those residues vital to maintaining the dimer interface. In fact, if SV3CP dimerisation could be proved essential for protease activity this would provide a further target for an inhibitory compound of SV3CP. Disruption of the dimer interface to prevent protease activity is proving a successful strategy with the HIV-1 protease; a similar approach could be applied to SV3CP. Though, of course the approach taken in treating the long term progression of HIV infection requiring a multi target approach is far removed from that required for treating SV infection. In the case of SV and other *Noro*viruses a single therapeutic agent that targets the 3C protease active site catalytic residues, such as MAPI, may be sufficient in arresting viral replication and therefore be capable of treating infection.

Reference list

1. Larsson, M. M., Rydell, G. E. P., Rodriguez-Diaz, J., Akerlind, B., Hutson, A. M., Estes, M. K., Larson, G., and Svensson, L. (2006) Antibody prevalence and titer to norovirus (genogroup II) correlate with secretor (FUT2) but not with ABO phenotype or Lewis (FUT3) genotype, *Journal of Infectious Diseases 194*, 1422-1427.

2. Hutson, A. E., Airaud, F., LePendu, J., Estes, M. K., and Atmar, R. L. (2005) Norwalk virus infection associates with secretor status genotyped from sera, *Journal of Medical Virology 77*, 455.

3. Marionneau, S., Airaud, F., Bovin, N. V., Le Pendu, J., and Ruvoen-Clouet, N. (2005) Influence of the combined ABO, FUT2, and FUT3 polymorphism on susceptibility to Norwalk virus attachment, *Journal of Infectious Diseases 192*, 1071-1077.

4. Tan, M. and Jiang, X. (2005) Norovirus and its histo-blood group antigen receptors: an answer to a historical puzzle, *Trends in Microbiology 13*, 285-293.

5. Lindesmith, L., Moe, C., LePendu, J., Frelinger, J. A., Treanor, J., and Baric, R. S. (2005) Cellular and humoral immunity following Snow Mountain virus challenge, *Journal of Virology 79*, 2900-2909.

6. Rockx, B. H. G., Vennema, H., Hoebe, C. J. P. A., Duizer, E., and Koopmans, M. P. G. (2005) Association of histo-blood group antigens and susceptibility to norovirus infections, *Journal of Infectious Diseases 191*, 749-754.

7. Hutson, A. M., Atmar, R. L., and Estes, M. K. (2004) Norovirus disease: changing epidemiology and host susceptibility factors, *Trends in Microbiology 12*, 279-287.

8. Huang, P. W., Farkas, T., Marionneau, S., Zhong, W. M., Ruvoen-Clouet, N., Morrow, A. L., Altaye, M., Pickering, L. K., Newburg, D. S., LePendu, J., and Jiang, X. (2003) Noroviruses bind to human ABO, Lewis, and secretor histo-blood group antigens: Identification of 4 distinct strain-specific patterns, *Journal of Infectious Diseases 188*, 19-31.

9. Lindesmith, L., Moe, C., Marionneau, S., Ruvoen, N., Jiang, X., Lindbland, L., Stewart, P., LePendu, J., and Baric, R. (2003) Human susceptibility and resistance to Norwalk virus infection, *Nature Medicine 9*, 548-553.

10. Farkas, T., Thornton, S. A., Wilton, N., Zhong, W., Altaye, M., and Jiang, X. (2003) Homologous versus heterologous immune responses to Norwalk-like viruses among crew members after acute gastroenteritis outbreaks on 2 US Navy vessels, *Journal of Infectious Diseases 187*, 187-193.

11. Harrington, P. R., Lindesmith, L., Yount, B., Moe, C. L., and Baric, R. S. (2002) Binding of Norwalk virus-like particles to ABH histo-blood group antigens is blocked by antisera from infected human volunteers or experimentally vaccinated mice, *Journal of Virology 76*, 12335-12343.

12. Marionneau, S., Ruvoen, N., Le Moullac-Vaidye, B., Clement, M., Cailleau-Thomas, A., Ruiz-Palacois, G., Huang, P. W., Jiang, X., and Le Pendu, J. (2002) Norwalk virus binds to histo-blood group antigens present on gastroduodenal epithelial cells of secretor individuals, *Gastroenterology 122*, 1967-1977.

13. Hutson, A. M., Atmar, R. L., Graham, D. Y., and Estes, M. K. (2002) Norwalk virus infection and disease is associated with ABO histo-blood group type, *Journal of Infectious Diseases 185*, 1335-1337.

14. Chang, J. G., Ko, Y. C., Lee, J. C. I., Chang, S. J., Liu, T. C., Shih, M. C., and Peng, C. T. (2002) Molecular analysis of mutations and polymorphisms of the Lewis secretor type alpha(1,2)-fucosyltransferase gene reveals that Taiwan aborigines are of Austronesian derivation, *Journal of Human Genetics 47*, 60-65.

15. Marionneau, S., Cailleau-Thomas, A., Rocher, J., Le Moullac-Vaidye, B., Ruvoen, N., Clement, M., and 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.

16. Ruvoen-Clouet, N., Ganiere, J. P., ndre-Fontaine, G., Blanchard, D., and Le Pendu, J. (2000) Binding of rabbit hemorrhagic disease virus to antigens of the ABH histo-blood group family, *Journal of Virology 74*, 11950-11954.

17. White, L. J., Ball, J. M., Hardy, M. E., Tanaka, T. N., Kitamoto, N., and Estes, M. K. (1996) Attachment and entry of recombinant Norwalk virus capsids to cultured human and animal cell lines, *Journal of Virology 70*, 6589-6597.

18. Rouquier, S., Lowe, J. B., Kelly, R. J., Fertitta, A. L., Lennon, G. G., and Giorgi, D. (1995) Molecular-Cloning of A Human Genomic Region Containing the H-Blood-Group Alpha(1,2)Fucosyltransferase Gene and 2 H-Locus-Related Dna Restriction Fragments - Isolation of A Candidate for the Human Secretor Blood-Group Locus, *Journal of Biological Chemistry 270*, 4632-4639.

19. Kelly, R. J., Rouquier, S., Giorgi, D., Lennon, G. G., and Lowe, J. B. (1995) Sequence and Expression of A Candidate for the Human Secretor Blood-Group Alpha(1,2)Fucosyltransferase Gene (Fut2) - Homozygosity for An Enzyme-Inactivating Nonsense Mutation Commonly Correlates with the Non-Secretor Phenotype, *Journal of Biological Chemistry 270*, 4640-4649.

20. Johnson, P. C., Mathewson, J. J., Dupont, H. L., and Greenberg, H. B. (1990) Multiple-Challenge Study of Host Susceptibility to Norwalk Gastroenteritis in United-States Adults, *Journal of Infectious Diseases 161*, 18-21.

21. Gary, G. W., Anderson, L. J., Keswick, B. H., Johnson, P. C., Dupont, H. L., Stine, S. E., and Bartlett, A. V. (1987) Norwalk Virus-Antigen and Antibody-Response in An Adult Volunteer Study, *Journal of Clinical Microbiology 25*, 2001-2003.

22. Koopman, J. S., Eckert, E. A., Greenberg, H. B., Strohm, B. C., Isaacson, R. E., and Monto, A. S. (1982) Norwalk Virus Enteric Illness Acquired by Swimming Exposure, *American Journal of Epidemiology 115*, 173-177.

23. Parrino, T. A., Schreiber, D. S., Trier, J. S., Kapikian, A. Z., and Blacklow, N. R. (1977) Clinical Immunity in Acute Gastroenteritis Caused by Norwalk Agent, *New England Journal of Medicine 297*, 86-89.

24. Nakamura, K., Someya, Y., Kumasaka, T., Ueno, G., Yamamoto, M., Sato, T., Takeda, N., Miyamura, T., and Tanaka, N. (2005) A norovirus protease structure provides insights into active and substrate binding site integrity, *Journal of Virology 79*, 13685-13693.

25. Zeitler, C. E., Estes, M. K., and Prasad, B. V. V. (2006) X-ray crystallographic structure of the Norwalk virus protease at 1.5-angstrom resolution, *Journal of Virology 80*, 5050-5058.

26. Mastrolorenzo, A., Rusconi, S., Scozzafava, A., and Supuran, C. T. (2006) Inhibitors of HIV-1 protease: 10 years after, *Expert Opinion on Therapeutic Patents 16*, 1067-1091.

27. Patick, A. K. and Potts, K. E. (1998) Protease inhibitors as antiviral agents, *Clinical Microbiology Reviews 11*, 614.

28. Kay, J. and Dunn, B. M. (1990) Viral Proteinases - Weakness in Strength, *Biochimica et Biophysica Acta 1048*, 1-18.

29. Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S., Merigan, T. C., Blaschke, T. F., Simpson, D., McLaren, C., Rooney, J., and Salgo, M. (1996) A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter, *New England Journal of Medicine 335*, 1081-1090.

30. Havlir, D. V., Eastman, S., Gamst, A., and Richman, D. D. (1996) Nevirapine-resistant human immunodeficiency virus: Kinetics of replication and estimated prevalence in untreated patients, *Journal of Virology 70*, 7894-7899.

31. Allaire, M., Chernaia, M. M., Malcolm, B. A., and James, M. N. G. (1994) Picornaviral 3C Cysteine Proteinases Have A Fold Similar to Chymotrypsin-Like Serine Proteinases, *Nature 369*, 72-76.

32. Peng, C., Ho, B. K., Chang, T. W., and Chang, N. T. (1989) Role of Human Immunodeficiency Virus Type-1-Specific Protease in Core Protein Maturation and Viral Infectivity, *Journal of Virology 63*, 2550-2556.

33. Martin, J. A. (1992) Recent Advances in the Design of HIV Proteinase-Inhibitors, *Antiviral Research 17*, 265-278.

34. Patick, A. K., Boritzki, T. J., and Bloom, L. A. (1997) Activities of the human immunodeficiency virus type 1 (HIV-1) protease inhibitor nelfinavir mesylate in combination with reverse transcriptase and protease inhibitors against acute HIV-1 infection in vitro, *Antimicrobial Agents and Chemotherapy 41*, 2159-2164.

35. Deeks, S. G., Smith, M., Holodniy, M., and Kahn, J. O. (1997) HIV-1 protease inhibitors - A review for clinicians, *Journal of the American Medical Association 277*, 145-153.

36. Mckinlay, M. A., Pevear, D. C., and Rossmann, M. G. (1992) Treatment of the Picornavirus Common Cold by Inhibitors of Viral Uncoating and Attachment, *Annual Review of Microbiology 46*, 635.

37. Lawson, M. A. and Semler, B. L. (1990) Picornavirus Protein Processing - Enzymes, Substrates, and Genetic-Regulation, *Current Topics in Microbiology and Immunology 161*, 49-87.

38. Makela, M. J., Puhakka, T., Ruuskanen, O., Leinonen, M., Saikku, P., Kimpimaki, M., Blomqvist, S., Hyypia, T., and Arstila, P. (1998) Viruses and bacteria in the etiology of the common cold, *Journal of Clinical Microbiology 36*, 539-542.

39. Matthews, D. A., Smith, W. W., Ferre, R. A., Condon, B., Budahazi, G., Sisson, W., Villafranca, J. E., Janson, C. A., Mcelroy, H. E., Gribskov, C. L., and Worland, S. (1994) Structure of human rhinovirus 3C protease reveals a trypsin-like polypeptide fold, RNA-binding site, and means for cleaving precursor Polyprotein, *Cell 77*, 761-771.

40. Dragovich, P. S., Webber, S. E., Babine, R. E., Fuhrman, S. A., Patick, A. K., Matthews, D. A., Lee, C. A., Reich, S. H., Prins, T. J., Marakovits, J. T., Littlefield, E. S., Zhou, R., Tikhe, J., Ford, C. E., Wallace, M. B., Meador, J. W., Ferre, R. A., Brown, E. L., Binford, S. L., Harr, J. E. V., DeLisle, D. M., and Worland, S. T. (1998) Structure-based design, synthesis, and biological evaluation of irreversible human rhinovirus 3C protease inhibitors. 1. Michael acceptor structure-activity studies, *Journal of Medicinal Chemistry 41*, 2806-2818.

41. Cordingley, M. G., Callahan, P. L., Sardana, V. V., Garsky, V. M., and Colonno, R. J. (1990) Substrate requirements of human rhinovirus 3C protease for peptide cleavage *in vitro*, *Journal of Biological Chemistry 265*, 9062-9065.

42. Dragovich, P. S., Webber, S. E., Babine, R. E., Fuhrman, S. A., Patick, A. K., Matthews, D. A., Reich, S. H., Marakovits, J. T., Prins, T. J., Zhou, R., Tikhe, J., Littlefield, E. S., Bleckman, T. M., Wallace, M. B., Little, T. L., Ford, C. E., Meador, J. W., Ferre, R. A., Brown, E. L., Binford, S. L., DeLisle, D. M., and Worland, S. T. (1998) Structure-based design, synthesis, and biological evaluation of irreversible human rhinovirus 3C protease inhibitors. 2. Peptide structure-activity studies, *Journal of Medicinal Chemistry 41*, 2819-2834.

43. Houghton, M., Weiner, A., Han, J., Kuo, G., and Choo, Q. L. (1991) Molecular-biology of the hepatitis-C viruses - Implications for diagnosis, development and control of viral disease, *Hepatology 14*, 381-388.

44. Zeuzem, S., Feinman, S. V., Rasenack, J., Heathcote, E. J., Lai, M. Y., Gane, E., O'Grady, J., Reichen, J., Diago, M., Lin, A., Hoffman, J., and Brunda, M. J. (2000) Peginterferon alfa-2a in patients with chronic hepatitis C, *New England Journal of Medicine 343*, 1666-1672.

45. Grakoui, A., Mccourt, D. W., Wychowski, C., Feinstone, S. M., and Rice, C. M. (1993) Characterization of the hepatitis-C virus-encoded serine proteinase - Determination of proteinase-dependent polyprotein cleavage sites, *Journal of Virology 67*, 2832-2843.

46. Grakoui, A., Mccourt, D. W., Wychowski, C., Feinstone, S. M., and Rice, C. M. (1993) A 2nd hepatitis-C virus-encoded proteinase, *Proceedings of the National Academy of Sciences of the United States of America 90*, 10583-10587.

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

48. Love, R. A., Parge, H. E., Wickersham, J. A., Hostomsky, Z., Habuka, N., Moomaw, E. W., Adachi, T., and Hostomska, Z. (1996) The crystal structure of hepatitis C virus NS3 proteinase reveals a trypsin-like fold and a structural zinc binding site, *Cell 87*, 331-342.

49. Koch, J. O. and Bartenschlager, R. (1997) Determinants of substrate specificity in the NS3 serine proteinase of the hepatitis C virus, *Virology 237*, 78-88.

50. Goudreau, N. and Llinas-Brunet, M. (2005) The therapeutic potential of NS3 protease inhibitors in HCV infection, *Expert Opinion on Investigational Drugs 14*, 1129-1144.

51. Reesink, H. W., Zeuzem, S., Weegink, C. J., Forestier, N., Van Vliet, A., De Rooij, J. V. D., McNair, L., Purdy, S., Kauffman, R., Alam, J., and Jansen, P. L. M. (2006) Rapid decline of viral RNA in hepatitis C patients treated with VX-950: A phase Ib, placebo-controlled, randomized study, *Gastroenterology 131*, 997-1002.

52. Darke, P. L., Jacobs, A. R., Waxman, L., and Kuo, L. C. (1999) Inhibition of hepatitis C virus NS2/3 processing by NS4A peptides - Implications for control of viral processing, *Journal of Biological Chemistry 274*, 34511-34514.

53. Koch, J. O. and Bartenschlager, R. (1997) Determinants of substrate specificity in the NS3 serine proteinase of the hepatitis C virus, *Virology 237*, 78-88.

54. Peng, C., Ho, B. K., Chang, T. W., and Chang, N. T. (1989) Role of Human Immunodeficiency Virus Type-1-Specific Protease in Core Protein Maturation and Viral Infectivity, *Journal of Virology 63*, 2550-2556.

55. Wyatt, R. G., Dolin, R., Blacklow, N. R., Dupont, H. L., Buscho, R. F., Thornhil, T. S., Kapikian, A. Z., and Chanock, R. M. (1974) Comparison of 3 agents of acute infectious nonbacterial gastroenteritis by cross-challenge in volunteers, *Journal of Infectious Diseases 129*, 709-714.

56. Liu, B. L., Clarke, I. N., and Lambden, P. R. (1996) Polyprotein processing in Southampton virus: Identification of 3C-like protease cleavage sites by in vitro mutagenesis, *Journal of Virology 70*, 2605-2610.

57. Belliot, G., Sosnovtsev, S. V., Chang, K. O., Babu, V., Uche, U., Arnold, J. J., Cameron, C. E., and Green, K. Y. (2005) Norovirus proteinase-polymerase and polymerase are both active forms of RNA-dependent RNA polymerase, *Journal of Virology 79*, 2393-2403.

58. Liu, B. L., Viljoen, G. J., Clarke, I. N., and Lambden, P. R. (1999) Identification of further proteolytic cleavage sites in the Southampton calicivirus polyprotein by expression of the viral protease in *E. coli*, *Journal of General Virology 80*, 291-296.

59. Kapikian, A. Z., Wyatt, R. G., Dolin, R., Thornhil, T. S., Kalica, A. R., and 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.

60. Dingle, K. E., Lambden, P. R., Caul, E. O., and Clarke, I. N. (1995) Human enteric Caliciviridae - the complete genome sequence and expression of virus-like particles from a genetic Group-Ii small round structured virus, *Journal of General Virology 76*, 2349-2355.

61. Jiang, X., Wang, M., Wang, K. N., and Estes, M. K. (1993) Sequence and genomic organization of Norwalk Virus, *Virology 195*, 51-61.

62. Lambden, P. R., Caul, E. O., Ashley, C. R., and Clarke, I. N. (1993) Sequence and genome organization of a human small round-structured (Norwalk-Like) virus, *Science 259*, 516-519.

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

64. Clarke, I. N. and Lambden, P. R. (2000) Organization and expression of calicivirus genes, *Journal of Infectious Diseases 181*, S309-S316.

65. Prasad, B. V. V., Matson, D. O., and Smith, A. W. (1994) 3-Dimensional Structure of Calicivirus, *Journal of Molecular Biology 240*, 256-264.

66. Lambden, P. R. and Clarke, I. N. (1995) Genome organization in the Caliciviridae, *Trends in Microbiology 3*, 261-265.

67. Wei, L., Huhn, J. S., Mory, A., Pathak, H. B., Sosnovtsev, S., Green, K. Y., and Cameron, C. E. (2001) Proteinase-polymerase precursor as the active form of feline calicivirus RNA-dependent RNA polymerase, *Journal of Virology 75*, 1211-1219.

68. Schultz-Cherry, S. (2005) Special focus section: Enteric viruses - Guest editorial: Update on enteric viruses, *Viral Immunology 18*, 2-3.

69. Cauchi, M. R., Doultree, J. C., Marshall, J. A., and Wright, P. J. (1996) Molecular characterization of Camberwell virus and sequence vacation in ORF3 of small round-structured (Norwalk-Like) viruses, *Journal of Medical Virology 49*, 70-76.

70. Seah, E. E. L., Marshall, J. A., and Wright, P. J. (1999) Open reading frame 1 of the Norwalk-like virus Camberwell: Completion of sequence and expression in mammalian cells, *Journal of Virology 73*, 10531-10535.

71. Someya, Y., Takeda, N., and Miyamura, T. (2000) Complete nucleotide sequence of the Chiba virus genome and functional expression of the 3C-like protease in Escherichia coli, *Virology 278*, 490-500.

72. Rinehart-Kim, J. E., Zhong, W. M., Jiang, X., Smith, A. W., and Matson, D. O. (1999) Complete nucleotide sequence and genomic organization of a primate calicivirus, Pan-1, *Archives of Virology 144*, 199-208.

73. Liu, B. L., Lambden, P. R., Gunther, H., Otto, P., Elschner, M., and Clarke, I. N. (1999) Molecular characterization of a bovine enteric calicivirus: Relationship to the Norwalk-like viruses, *Journal of Virology 73*, 819-825.

74. Love, D. N. and Sabine, M. (1975) Electron-microscopic observation of feline kidney-cells infected with a feline Calicivirus, *Archives of Virology 48*, 213-228.

75. Peeters, J. E. (1990) Viral Hemorrhagic-Disease, a new threat for the rabbit production, *Vlaams Diergeneeskundig Tijdschrift 59*, 208-212.

76. Maillard, J. Y. (2001) Virus susceptibility to biocides: an understanding, *Reviews in Medical Microbiology 12*, 63-74.

77. Matsui, S. M., Kim, J. P., Greenberg, H. B., Su, W. C., Sun, Q. M., Johnson, P. C., Dupont, H. L., Oshiro, L. S., and Reyes, G. R. (1991) The isolation and characterization of a Norwalk virus-specific cDNA, *Journal of Clinical Investigation 87*, 1456-1461.

78. Deleon, R., Matsui, S. M., Baric, R. S., Herrmann, J. E., Blacklow, N. R., Greenberg, H. B., and Sobsey, M. D. (1992) Detection of Norwalk Virus in stool specimens by reverse transcriptase-polymerase chain-reaction and nonradioactive oligoprobes, *Journal of Clinical Microbiology 30*, 3151-3157.

79. Hardy, M. E. and Estes, M. K. (1996) Completion of the Norwalk virus genome sequence, *Virus Genes 12*, 287-290.

80. Dingle, K. E., Lambden, P. R., Caul, E. O., and Clarke, I. N. (1995) Human enteric Caliciviridae - the complete genome sequence and expression of virus-like particles from a genetic Group-Ii small round structured virus, *Journal of General Virology 76*, 2349-2355.

81. Hamre, D., Bernstein, J., and Donovick, R. (1950) Activity of *para*-aminobenzaldehyde, 3-thiosemicarbazone on vaccinia virus in the chick embryo and in the mouse, *Proceedings of the Society of Experimental Biological Medicine 73*, 275-278.

82. Furman, P. A., Mcguirt, P. V., Keller, P. M., Fyfe, J. A., and Elion, G. B. (1980) Inhibition by Acyclovir of cell-Growth and DNA-Synthesis of cells biochemically transformed with Herpes virus genetic information, *Virology 102*, 420-430.

83. Rooke, R., Tremblay, M., Lalande, C., Parniak, M. A., and Wainberg, M. A. (1991) Isolation of HIV-1 strains resistant to Azt - limitations in the use of nucleoside analogs in the treatment of AIDS, *Medicine Sciences 7*, 118-126.

84. Gianotti, N., Soria, A., and Lazzarin, A. (2007) Antiviral activity and clinical efficacy of atazanavir in HIV-1-infected patients: a review, *New Microbiologica 30*, 79-88.

85. Mastrolorenzo, A., Rusconi, S., Scozzafava, A., and Supuran, C. T. (2006) Inhibitors of HIV-1 protease: 10 years after, *Expert Opinion on Therapeutic Patents 16*, 1067-1091.

86. Palumbo, E. (2007) PEG-interferon alfa-2b for acute hepatitis C: A review, *Mini-Reviews in Medicinal Chemistry 7*, 839-843.

87. Herrine, S. K., Rossi, S., and Navarro, V. J. (2006) Management of patients with chronic hepatitis C infection, *Clinical and Experimental Medicine 6*, 20-26.

88. Weigand, K., Stremmel, W., and Encke, J. (2007) Treatment of hepatitis C virus infection, *World Journal of Gastroenterology 13*, 1897-1905.

89. Dragovich, P. S., Prins, T. J., Zhou, R., Johnson, T. O., Hua, Y., Luu, H. T., Sakata, S. K., Brown, E. L., Maldonado, F. C., Tuntland, T., Lee, C. A., Fuhrman, S. A., Zalman, L. S., Patick, A. K., Matthews, D. A., Wu, E. Y., Guo, M., Borer, B. C., Nayyar, N. K., Moran, T., Chen, L. J., Rejto, P. A., Rose, P. W., Guzman, M. C., Dovalsantos, E. Z., Lee, S., Mcgee, K., Mohajeri, M., Liese, A., Tao, J. H., Kosa, M. B., Liu, B., Batugo, M. R., Gleeson, J. P. R., Wu, Z. P., Liu, J., Meador, J. W., and Ferre, R. A. (2003) Structure-based design, synthesis, and biological evaluation of irreversible human rhinovirus 3C protease inhibitors. 8. Pharmacological optimization of orally bioavailable 2-pyridone-containing peptidomimetics, *Journal of Medicinal Chemistry 46*, 4572-4585.

90. Garman, E. and Nave, C. (2002) Radiation damage to crystalline biological molecules: current view, *Journal of Synchrotron Radiation 9*, 327-328.

91. Weik, M., Ravelli, R. B. G., Kryger, G., McSweeney, S., Raves, M. L., Harel, M., Gros, P., Silman, I., Kroon, J., and Sussman, J. L. (2000) Specific chemical and structural damage to proteins produced by synchrotron radiation, *Proceedings of the National Academy of Sciences of the United States of America 97*, 623-628.

92. Chayen, N. E. (1998) Comparative studies of protein crystallization by vapour-diffusion and microbatch techniques, *Acta Crystallographica Section D-Biological Crystallography 54*, 8-15.

93. Feher, G. and Kam, Z. (1985) Nucleation and growth of protein crystals - general-principles and assays, *Methods in Enzymology 114*, 77-112.

94. Drenth, J. and Haas, C. (1992) Protein crystals and their stability, *Journal of Crystal Growth 122*, 107-109.

95. Kam, Z., Shore, H. B., and Feher, G. (1978) Crystallization of proteins, *Journal of Molecular Biology 123*, 539-555.

96. Guilloteau, J. P., Rieskautt, M. M., and Ducruix, A. F. (1992) Variation of Lysozyme Solubility As a function of temperature in the presence of organic and inorganic salts, *Journal of Crystal Growth 122*, 223-230.

97. Ducruix, A. F. and Giege, R. (1992) *Crystallisation of nucleic acids and proteins: A practical approach* Oxford University Press.

98. Aguilar, C. F., Newman, M. P., Aparicio, J. S., Cooper, J. B., Tickle, I. J., and Blundell, T. L. (1993) The use of protein homologs in the rotation function, *Acta Crystallographica Section A 49*, 306-315.

99. Brunger, A. T. (1992) Free R-Value - A novel statistical quantity for assessing the accuracy of crystal-structures, *Nature 355*, 472-475.

100. Brunger, A. T. (1990) Extension of Molecular Replacement - A new search strategy based on Patterson Correlation Refinement, *Acta Crystallographica Section A 46*, 46-57.

101. Brunger, A. T. (1997) X-ray crystallography and NMR reveal complementary views of structure and dynamics, *Nature Structural Biology 4*, 862-865.

102. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977) Protein Data Bank - Computer-based archival file for Macromolecular structures, *Journal of Molecular Biology 112*, 535-542.

103. Garman, E. F. and Schneider, T. R. (1997) Macromolecular cryocrystallography, *Journal of Applied Crystallography 30*, 211-237.

104. Ravelli, R. B. G. and Garman, E. F. (2006) Radiation damage in macromolecular cryocrystallography, *Current Opinion in Structural Biology 16*, 624-629.

105. Rhodes, G. (2000) *Crystallography: Made crystal clear* Academic Press.

106. Blow, D.M. (2002) *Outline of Crystallography for Biologists* Oxford University Press.

107. Brady, L., Cooper, J. B., Garman, E., McCoy, A., Noble, M., Emsley, P., Keep, N., and Artymiuk, P. (2004) BCA Summer School in Protein Crystallography.

108. Nanev, C. N. (2007) On the slow kinetics of protein crystallization, *Crystal Growth & Design 7*, 1533-1540.

109. Blundell, T. L. and Johnson, L. N. (1976) *Protein Crystallography* Academic Press.

110. Blow, D. M. (2003) How Bijvoet made the difference: The growing power of anomalous scattering, *Methods in Enzymology 372*, 3-22.

111. Mcpherson, A. (1990) Current approaches to macromolecular crystallization, *European Journal of Biochemistry 189*, 1-23.

112. Giacovazzo, C., Monaco, H. L., Viterbo, D., Scordari, F., Gilli, G., Zanotti, G., and Catti, M. (1992) *Fundamentals of Crystallography* Oxford University Press.

113. Drenth, J. (1994) *Principles of Protein X-ray Crystallography* Springer-Verlag.

114. Mitchell, E., Kuhn, P., and Garman, E. (1999) Demystifying the synchrotron trip: a first time users guide, *Structure 7*, 111-121.

115. Bailey, S. (1994) The Ccp4 Suite - Programs for protein crystallography, *Acta Crystallographica Section D-Biological Crystallography 50*, 760-763.

116. Leslie, A. G. W. (2006) The integration of macromolecular diffraction data, *Acta Crystallographica Section D-Biological Crystallography 62*, 48-57.

117. Leslie, A. G. W. (1997) MOSFLM user guide, MRC Laboratory of Molecular Biology, Cambridge.

118. Terwilliger, T. C., Kim, S. H., and Eisenberg, D. (1987) Generalized method of determining heavy-atom positions using the difference Patterson Function, *Acta Crystallographica Section A 43*, 1-5.

119. Terwilliger, T. C. and Eisenberg, D. (1983) Unbiased 3-dimensional refinement of heavy-atom parameters by correlation of origin-removed Patterson Functions, *Acta Crystallographica Section A 39*, 813-817.

120. Terwilliger, T. C. and Eisenberg, D. (1987) Isomorphous Replacement - Effects of errors on the phase probability-distribution, *Acta Crystallographica Section A 43*, 6-13.

121. Terwilliger, T. C. (1994) Mad Phasing - Bayesian Estimates of F-A, *Acta Crystallographica Section D-Biological Crystallography 50*, 11-16.

122. Terwilliger, T. C. (1994) Mad Phasing - Treatment of dispersive differences as Isomorphous Replacement information, *Acta Crystallographica Section D-Biological Crystallography 50*, 17-23.

123. Terwilliger, T. C. and Berendzen, J. (1997) Bayesian correlated MAD phasing, *Acta Crystallographica Section D-Biological Crystallography 53*, 571-579.

124. Brunger, A. T., Adams, P. D., Clore, G. M., Delano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination, *Acta Crystallographica Section D-Biological Crystallography 54*, 905-921.

125. Ogata, C. M. (1998) MAD phasing grows up, *Nature Structural Biology 5*, 638-640.

126. Hendrickson, W. A. and Ogata, C. M. (1997) Phase determination from multiwavelength anomalous diffraction measurements, *Methods in Enzymology 276,* 494-523.

127. Walsh, M. A., Evans, G., Sanishvili, R., Dementieva, I., and Joachimiak, A. (1999) MAD data collection - current trends, *Acta Crystallographica Section D-Biological Crystallography 55*, 1726-1732.

128. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank, *Nucleic Acids Research 28*, 235-242.

129. Helliwell, J. R. (1992) Synchrotron radiation instrumentation and macromolecular crystallography, *Review of Scientific Instruments 63*, 1628-1629.

130. Winn, M. D., Murshudov, G. N., and Papiz, M. Z. (2003) Macromolecular TLS refinement in REFMAC at moderate resolutions, *Methods in Enzymology 374,* 300-321.

131. Sheldrick, G. M. and Schneider, T. R. (1997) SHELXL: High-resolution refinement, *Methods in Enzymology 277,* 319-343.

132. Perrakis, A., Harkiolaki, M., Wilson, K. S., and Lamzin, V. S. (2001) ARP/wARP and molecular replacement, *Acta Crystallographica Section D-Biological Crystallography 57*, 1445-1450.

133. Bergfors, T. (2003) Seeds to crystals, *Journal of Structural Biology 142*, 66-76.

134. Blow, D. M. (1985) Introduction to rotation and translation functions, *Proceedings of the Daresbury Study Weekend.*

135. Garman, E. (2003) 'Cool' crystals: macromolecular cryocrystallography and radiation damage, *Current Opinion in Structural Biology 13*, 545-551.

136. Garman, E. (1999) Cool data: quantity AND quality, *Acta Crystallographica Section D-Biological Crystallography 55*, 1641-1653.

137. Garman, E. and Murray, J. W. (2003) Heavy-atom derivatization, *Acta Crystallographica Section D-Biological Crystallography 59*, 1903-1913.

138. Hendrickson, W. A. (2000) Synchrotron crystallography, *Trends in Biochemical Sciences 25*, 637-643.

139. Patterson, A. L. (1934) A Fourier series method for the determination of the components of interatomic distances in crytals, *Physics Review 46*, 372-376.

140. Delano, W. L. (2002) The PyMOL Molecular Graphics System, DeLano Scientific, Palo Alto, CA, USA.

141. Cruickshank, D. W. J. (1961) *Computing Methods and the Phase Problem* Pergamon Press.

142. Terwilliger, T. C. and Berendzen, J. (1999) Automated MAD and MIR structure solution, *Acta Crystallographica Section D-Biological Crystallography 55*, 849-861.

143. Terwilliger, T. C. (1999) Reciprocal-space solvent flattening, *Acta Crystallographica Section D-Biological Crystallography 55*, 1863-1871.

144. Cowtan, K. D. and Zhang, K. Y. J. (1999) Density modification for macromolecular phase improvement, *Progress in Biophysics & Molecular Biology 72*, 245-270.

145. Levitt, D. G. (2001) A new software routine that automates the fitting of protein X-ray crystallographic electron-density maps, *Acta Crystallographica Section D-Biological Crystallography 57*, 1013-1019.

146. Laskowski, R. A., Macarthur, M. W., Moss, D. S., and Thornton, J. M. (1993) Procheck - A Program to Check the Stereochemical Quality of Protein Structures, *Journal of Applied Crystallography 26*, 283-291.

147. Lovell, S. C., Davis, I. W., Adrendall, W. B., de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S., and Richardson, D. C. (2003) Structure validation by C alpha geometry: phi,psi and C beta deviation, *Proteins-Structure Function and Genetics 50*, 437-450.

148. Merrifield, R. B. (2007) Solid phase peptide synthesis.I. The synthesis of a tetrapeptide, *Journal of the American Chemical Society 85*, 2149-2154.

149. Atherton, E., Fox, H., Harkiss, D., Logan, C. J., Sheppard, R. C., and Williams, B. J. (1978) Mild procedure for solid-phase peptide-synthesis - Use of Fluorenylmethoxycarbonylamino-Acids, *Journal of the Chemical Society-Chemical Communications* 537-539.

150. Fields, G. B. and Noble, R. L. (1990) Solid-phase peptide-synthesis utilizing 9-fluorenylmethoxycarbonyl amino-acids, *International Journal of Peptide and Protein Research 35*, 161-214.

151. Chan, W. C. and White, P. D. (2000) *Fmoc Solid Phase Peptide Synthesis* Oxford University Press.

152. Whitmore, A. J., Daniel, R. M., and Petach, H. H. (1995) A general method for the synthesis of peptidyl substrates for proteolytic enzymes, *Tetrahedron Letters 36*, 475-476.

153. Kaspari, A., Schierhorn, A., and Schutkowski, M. (1996) Solid-phase synthesis of peptide-4-nitroanilides, *International Journal of Peptide and Protein Research 48*, 486-494.

154. Rink, H. (1987) Solid-phase synthesis of protected peptide fragments using a trialkoxy-diphenyl-methylester resin, *Tetrahedron Letters 28*, 3787-3790.

155. Bernatowicz, M. S., Daniels, S. B., and Koster, H. (1989) A comparison of acid labile linkage agents for the synthesis of peptide C-terminal amides, *Tetrahedron Letters 30*, 4645-4648.

156. Broadbridge, R. (2002). Personal communication.

157. Liu, S. and Hanzlik, R. P. (1992) Structure-activity-relationships for inhibition of papain by peptide Michael Acceptors, *Journal of Medicinal Chemistry 35*, 1067-1075.

158. Govardhan, C. P. and Abeles, R. H. (1996) Inactivation of cysteine proteases, *Archives of Biochemistry and Biophysics 330*, 110-114.

159. Dragovich, P. S., Webber, S. E., Babine, R. E., Fuhrman, S. A., Patick, A. K., Matthews, D. A., Reich, S. H., Prins, T. J., Marakovits, J. T., Littlefield, E. S., Zhou, R., Tikhe, J., Ford, C. E., Wallace, M. B., Bleckman, T. M., Meador, J. W., Ferre, R. A., Brown, E. L., Binford, S. L., DeLisle, D. M., and Worland, S. T. (1998) Structure-based design of irreversible human rhinovirus 3C protease inhibitors, *Abstracts of Papers of the American Chemical Society 215*, 863.

160. Nahm, S. and Weinreb, S. M. (1981) N-methoxy-N-methylamides as effective acylating agents, *Tetrahedron Letters 22*, 3815-3818.

161. Furlong, J., Meighan, M., Conner, J., Murray, J., and Clements, J. B. (1992) Methods for improved protein expression using pET vectors, *Nucleic Acids Research 20*, 4668.

162. Ramakrishnan, V. and Graziano, V. (2002) Recombinant expression of selenomethionine derivative proteins in *E. coli.* http://alf1.mrc-lmb.ac.uk/~ramak/madms/segrowth.html

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

164. Mosimann, S. C., Cherney, M. M., Sia, S., Plotch, S., and James, M. N. G. (1997) Refined x-ray crystallographic structure of the poliovirus 3C gene product, *Journal of Molecular Biology 273*, 1032-1047.

165. Matthews, D. A., Dragovich, P. S., Webber, S. E., Fuhrman, S. A., Patick, A. K., Zalman, L. S., Hendrickson, T. F., Love, R. A., Prins, T. J., Marakovits, J. T., Zhou, R., Tikhe, J., Ford, C. E., Meador, J. W., Ferre, R. A., Brown, E. L., Binford, S. L., Brothers, M. A., DeLisle, D. M., and Worland, S. T. (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 of the United States of America 96*, 11000-11007.

166. Barrette-Ng, I. H., Ng, K. K. S., Mark, B. L., van Aken, D., Cherney, M. M., Garen, C., Kolodenko, Y., Gorbalenya, A. E., Snijder, E. J., and James, M. N. G. (2002) Structure of arterivirus nsp4 - The smallest chymotrypsin-like proteinase with an alpha/beta C-terminal extension and alternate conformations of the oxyanion hole, *Journal of Biological Chemistry 277*, 39960-39966.

167. Bergmann, E. M., Mosimann, S. C., Chernaia, M. M., Malcolm, B. A., and James, M. N. G. (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.

168. Bergmann, E. M., Cherney, M. M., Mckendrick, J., Frormann, S., Luo, C., Malcolm, B. A., Vederas, J. C., and James, M. N. G. (1999) Crystal structure of an inhibitor complex of the 3C proteinase from hepatitis a virus (HAV) and implications for the polyprotein processing in HAV, *Virology 265*, 153-163.

169. Zeitler, C. E., Estes, M. K., and Prasad, B. V. V. (2006) X-ray crystallographic structure of the Norwalk virus protease at 1.5-angstrom resolution, *Journal of Virology 80*, 5050-5058.

170. Nakamura, K., Someya, Y., Kumasaka, T., Ueno, G., Yamamoto, M., Sato, T., Takeda, N., Miyamura, T., and Tanaka, N. (2005) A norovirus protease structure provides insights into active and substrate binding site integrity, *Journal of Virology 79*, 13685-13693.

171. Someya, Y., Takeda, N., and Miyamura, T. (2002) Identification of active-site amino acid residues in the Chiba virus 3C-like protease, *Journal of Virology 76*, 5949-5958.