University of Southampton

Quasi Laue Neutron and Atomic resolution X-ray Diffraction of Endothiapepsin

A Thesis presented by

Leighton Coates

To the University of Southampton, faculty of science for the degree of Doctor of Philosophy.

School of Biological Sciences
University of Southampton
Bassett Crescent East
Southampton
S016 7PX
UK

August 2002

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SCIENCE-DIVISION OF BIOCHEMISTRY

DOCTOR OF PHILOSOPHY

QUASI LAUE NEUTRON AND ATOMIC X-RAY DIFFRACTION OF ENDOTHIAPEPSIN

by Leighton Coates

Endothiapepsin is derived from the fungus *Endothia parasitica* and is a member of the aspartic proteinase class of enzymes. This class of enzyme is comprised of two structurally similar lobes; each lobe contributes an aspartic acid residue to form a catalytic dyad that acts to cleave the substrate peptide bond. Knowledge of the protonation states of these aspartates in the tetrahedral intermediate state would determine the catalytic mechanism by which the enzyme operates. The three dimensional structure of endothiapepsin bound to the transition state analogue inhibitor H261 has been solved to high resolution (2.1Å) using quasi Laue neutron diffraction. At the time of writing this is the largest protein structure determined at high resolution using neutron diffraction. The position of deuterium atoms in the active site indicates that the outer oxygen of Asp 215 and the inhibitory hydroxyl group are protonated in the transition state analogue complex. The three dimensional structures of endothiapepsin bound to five transition state analogue inhibitors (H189, H256, CP-80,794, PD-129,541 and PD-130,328) have also been solved using X-rays to atomic resolution allowing full anisotropic modelling of each complex. The structure of endothiapepsin complexed with the *gem*-diol based inhibitor PD-135,040 has also been solved to a resolution of 1.6 Å. The active sites of the six structures have been studied with a view to studying the catalytic mechanism of the aspartic proteinases by locating the active site protons by carboxyl bond length differences and electron density analysis. In the CP-80,794 structure there is excellent electron density for the hydrogen on the inhibitory statine hydroxyl group which forms a hydrogen bond with the inner oxygen of Asp 32. A number of short hydrogen bonds that may have a role in catalysis (~2.6 Å) have been identified within the active site in each structure; the presence of these bonds has been confirmed using NMR techniques.

Table of Contents

Table of Figures

List of tables

Ozymandias

I met a traveller from an antique land
Who said: Two vast and trunkless legs of stone
Stand in the desert. Near them, on the sand,
Half sunk, a shattered visage lies, whose frown,
And wrinkled lip, and sneer of cold command,
Tell that its sculptor well those passions read
Which yet survive, stamped on these lifeless things,
The hand that mocked them, and the heart that fed;
And on the pedestal these words appear:
"My name is Ozymandias, king of kings:
Look on my works, ye Mighty, and despair!"
Nothing beside remains. Round the decay
Of that colossal wreck, boundless and bare
The lone and level sands stretch far away.

Percy Bysshe Shelley

Acknowledgements

This thesis is dedicated to my father Robert Coates and to the memory of my mother Denise Coates.

I would like to thank Jon Cooper for training and instruction during my PhD. On a more general level I would like to thank the following past and present members of the crystallography group for help and friendship Peter Erskine, Fiyaz Mohammed, Darren Thompson, Royston Gill, Mark Montgomery, Paul Williams, Sanjay Mall, Alan Purvis, Terry Robinson and Steve Wood.

Abbreviations

ϕ, phi

$\otimes$, convolution

μ, linear absorption coefficient

μ', altered linear absorption coefficient

µl, micro litre

ω, Angular velocity

Å, Ångstrom 10^{-10} metres

ACE, angiotensin converting enzyme

ADP, anisotropic displacement parameter

AIDS, acquired immunodeficiency syndrome

ASP, Aspartic acid

b, scattering length barns

b+, coherent scattering length barns

b-, incoherent scattering length barns

barn, 10^{-28} M^2

$\mathbf{B_{iso}}$, isotropic displacement parameter in $Å^2$

CCD, charge coupled device

CCP4, collaborative computing project number 4

CGLS, conjugate gradient least squares

ΔH, change in enthalpy

$\mathbf{D_{hkl}}$, interplanar spacing

$\mathbf{D_{min}}$, minimum interplanar spacing

ESD, Estimated standard deviation

ESRF, European synchrotron radiation facility

$\mathbf{F_c}$, calculated structure factor

$\mathbf{F_{hkl}}$, structure factor

$\mathbf{F_o}$, observed structure factor

***gem*-diol,** geminal-diol

h, Planks constant

HIV, human immunodeficiency virus

I, intensity

ILL, Insitute Laue Langevin

$\mathbf{I_{tot}}$, total intensity

$\mathbf{I_{bkg}}$, background intensity

K, Kelvin

$\mathbf{k_{cat}}$, rate constant or turnover number

KDa, kilo dalton

$\mathbf{K_i}$, inhibition constant

$\mathbf{K_m}$, Michealis-Menton constant

KSI, ketosteroid isomerase

λ, wavelength

λ_{max}, maximum wavelength

λ_{min}, minimum wavelength

LADI, Laue diffractometer

LBHB, low barrier hydrogen bond

m, mass

meV, micro electron volt

MeV, milli electron volt

mg, milli gram

ml, milli litre

mM, milli moles per litre

n.a., numerical aperture

NCS, non crystallographic symmetry

NIP, neutron image plate

nM, nano moles per litre

NMR, nuclear magnetic resonance

PDB, protein databank

pH, $-\log_{10}$ hydrogen ion concentration

pK, dissociation constant

$\mathbf{pK_a}$, acid dissociation constant

ppm, parts per million

r, 3d spatial coordinate

RLP, reciprocal lattice point

RMS, root mean square

RMSD, root mean square deviation

σ(I), error in associated intensity measurement

S, scattering vector

SDS, sodium dodecyl sulphate

θ, scattering angle

T, temperature

U, mean square displacement in $Å^2$

v, velocity

Z, atomic number

Chapter 1

Introduction to the aspartic proteinases

Aspartic proteinases are a widely distributed class of enzymes found in fungi, plants and vertebrates as well as being found in the HIV retrovirus where the enzyme is essential for maturation of the virus particle. They play a major role in a number of diseases and pathogen infections. For a recent review of the role of aspartic proteinases in disease see Cooper (2001). The best-known aspartic proteinase is pepsin, which is involved in digestion in the stomach. Endothiapepsin is an aspartic proteinase associated with the chestnut blight fungus. The fungus secretes the proteinase where it has a role in digesting the growth medium. Aspartic proteinases comprise of a large family of enzymes that are involved in a number of important physiological and pathological processes. Aspartic proteinases cut the target peptide chain via two catalytic aspartate residues, which are held in close proximity to each other via a complex arrangement of hydrogen bonds (Bailey and Cooper, 1994). A single water molecule, which forms tight hydrogen bonds to both aspartic carboxyl groups is presumed to take part in the catalytic mechanism (Pearl and Blundell, 1984). Like most aspartic proteinases endothiapepsin has an optimal acidic pH (5.5). It cleaves protein substrates with a similar specificity to that of porcine pepsin a, as it prefers hydrophobic residues at each side of the cleavage site. Williams *et al.*, (1972) analysed the cleavage sites in the oxidised B-chain of insulin and found the rates of cleavage for endothiapepsin to be as follows: Phe-Phe > Tyr-Leu > Gln-His >>> Leu Val > Asn-Gln. Like most aspartic proteinases endothiapepsin is strongly inhibited by the microbial hexapeptide pepstatin-A, which contains the unusual amino acid statine (Bailey *et al.*, 1993; Cooper *et al.*, 1989). The statine residue contains a main chain -CHOH-CH_2- group which is thought to act as an analogue of the putative tetrahedral intermediate (-$(OH)_2$-NH-) produced during catalysis. The statine residue was found in a potent aspartic proteinase inhibitor derived from several species of the actinomyces (moulds) by Umezawa *et al* (1970). The inhibitor was called pepstatin due to its effectiveness in inhibiting pepsin.

The role of aspartic proteinases in disease

Hypertension

Hypertension is a major risk factor for coronary heart disease and stroke particularly in the western world; it frequently causes damage to the arterial blood vessels, the eyes and kidneys. Prolonged hypertension also causes enlargement of the heart and may ultimately lead to heart failure. The occurrence of hypertension increases with age with more than half of the entire population in the western world over the age of 60 having high blood pressure. A general decline in death rates from coronary heart disease and stroke in recent decades can be attributed in part to improvements in the treatment and control of hypertension. One of the key mediators in primary hypertension is the plasma octapeptide angiotensin II (AII) that plays a major role in hypertension by causing vasoconstriction and stimulating aldosterone release, thereby increasing blood volume by the action of aldosterone on the kidneys. Angiotensin II is produced from a proteolytic cascade known as the renin-angiotensin system in which the aspartic proteinase renin catalyses the rate limiting cleavage of angiotensinogen, produced by the liver, to yield the decapeptide angiotensin I (AI). The subsequent removal of the carboxy-terminal dipeptide from AI by angiotensin converting enzyme (ACE), yields AII. ACE is the target for a number of widely prescribed drugs which are effective for treating hypertension, hyperaldosteronism and congestive heart failure (Ondetti *et al* 1982). The development of potent low molecular weight orally active ACE inhibitors from natural and synthetic metalloproteinase inhibitors has been rapid due, in part, to the relative lack of specificity of the enzyme. The degradation of bradykinin and other members of the kinin family by ACE may be responsible for some of the side effects of ACE inhibitors. In contrast, the aspartic proteinase renin cleaves only its natural substrate or very close analogues and although inhibition of an enzyme more specific than ACE may be desirable for reducing side-effects *in vivo*, the selectivity of renin meant that during the early stages of drug development, potent inhibition required the use of large peptide-based

compounds. These were often poorly absorbed and susceptible to gastric proteolysis and biliary excretion. Nevertheless, the commercial and clinical success of ACE inhibitors fuelled interest in the search for therapeutic renin drugs during the 1980's. Although renin remains an excellent potential drug target, the problems of poor oral bioavailability and rapid excretion of peptide drugs have been largely insurmountable to the extent that virtually all renin programmes in the pharmaceutical industry have been closed. Nevertheless, many of the inhibitors developed during this period proved to be successful lead compounds in the search for inhibitors of HIV proteinase.

HIV and AIDS

The current world AIDS epidemic is affecting developing countries hardest, particularly sub-saharan Africa. It is a major health concern that has now become a global tragedy with treatment affordable only for a small percentage of infected people (Piot *et al*). It is estimated that 36 million people are currently living with HIV/AIDS; 22 million men, women, and children have already died, and 15,000 new infections occur each day. Following the initial infection by the HIV virus, it can be up to five years before the onset of any symptoms of AIDS are observed. This time period is dependent on many factors including the number of T-helper cells in the immune system that the virus is able to infect. In addition the age and the general health condition of the person also plays a role. The progressive weakening of the immune system eventually makes the victim more and more susceptible to opportunistic infections. As the immune system becomes increasingly compromised, it is not able to fight off more serious infections that a normal intact immune system could suppress. Some of these infections can be life threatening to a person with AIDS and are usually the final cause of death.

The genome of the AIDS retrovirus is encoded in 10,000 bases of RNA (Tomasselli *et al* 2000) and is highly prone to mutation. The genome is composed of 3 reading frames: *gag*, *pol* and *env* that code for several proteins that are essential for virus assembly and replication. The *gag* gene encodes proteins that

make up the viral core, *pol:* encodes the enzyme reverse transcriptase, and *env:* encodes proteins that make up the viral envelope. The enzyme reverse transcriptase that is present within the virus particle copies the retroviral RNA sequence into a single-stranded DNA molecule when infection of a host cell has occurred. A complementary strand of DNA is then generated and the resulting double-stranded DNA copy of the retroviral genome integrates into the host cell's DNA. The integrated proviral genome can persist in this state for many years before any AIDS symptoms finally develop due to destruction of helper T-cells. Expression of the retroviral genes by the host cell leads to the formation of new copies of the virus.

The *gag*, *pol* and *env* reading frames are expressed as poly-proteins that eventually have to be separated in order for each of the individual protein molecules to perform its function. This cleavage is performed by HIV proteinase a member of the aspartic proteinase family and is encoded by part of the *pol* gene. The proteolytic maturation of virus particles continues after they have 'budded' from the host cell. It has been observed that mutant viruses containing a catalytically inactive proteinase fail to mature and hence are not infectious. Thus inhibitors of the proteinase are widely used in multi-drug therapy and improved compounds are much sought after as drugs for treatment of HIV infection.

Alzheimer's Disease

Alzheimer's disease (AD) is the most common cause of dementia in older people. The disease is named after the German clinician (A. Alzheimer) who, in 1906, discovered changes in the brain tissue of a woman who had died of an unusual mental illness. The progression of Alzheimer's disease involves the destruction of cells that control memory especially in the hippocampus and other related structures of the brain (Fassbender *et al* 2001). Subsequently the speech and reasoning of the patient deteriorates as the cerebral cortex becomes affected. Drastic personality changes occur along with emotional outbursts and other disturbed behaviour. Eventually many other areas of the brain become affected

with all regions shrinking until the patient becomes completely helpless and eventually dies. Alzheimer's disease can occur sporadically or as an inherited trait with the majority of patients developing the illness beyond the age of 65 although 10 % of victims develop the disease before reaching this age.

Alzheimer's disease is caused by the formation of amyloid plaques in the brain and neurofibrillary tangles in the nerves and is one of a number of the so-called amyloid diseases. An amyloid plaque is an abnormal cluster of dead and dying nerve cells, other brain cells, and amyloid protein fragments. In these deposits the proteins are often aberrantly folded into structures different to the ones that they would normally adopt. The plaques have been shown by electron microscopy to contain protein fibrils of indefinite length. X-ray fibre diffraction suggests that these fibrils consist of polypeptide in the twisted β-pleated sheet conformation (Serpell *et al* 2000). The fibrils within neuritic plaques are formed from a protein known as amyloid β-protein (Aβ) that is produced by cleavage of a transmembrane glycoprotein of unknown function which is referred to as amyloid precursor protein (APP) (Hong *et al* 2000). There are 8 members in the APP family which are generated by alternative splicing of 3 exons. The amyloid β-protein is 40-42 residues in length and is generated by cleavage of the large extracellular domain of the APP precursor at two sites one of which is close to the transmembrane region and the other is within it. This amyloidogenic processing is catalysed by proteinases called β- and γ-secretase. It is the γ- secretase enzyme which cleaves APP within its transmembrane domain and the β-secretase cleaves APP on the lumenal side of the membrane. An alternative, non-amyloidogenic pathway of APP processing which is catalysed by α-secretase involves cleavage within the amyloid β-peptide region of the precursor. Both pathways are operative in neurones; the only cells in the brain that constitutively express APP although other cells express it on activation. For each secretase there are several sites of cleavage in APP which are very close together. In familial Alzheimer's disease, there are several known mutations which occur at the α-, β- and γ-secretase cleavage sites. These may affect the pattern of APP processing and lead to increased production of the fibril-

forming amyloid β-peptide. Recently it has been shown that β-secretase is an aspartic proteinase and there is major interest in designing drugs which could inhibit this enzyme as a potential therapy for Alzheimer's disease (Hong *et al* 2000).

Breast Cancer

Breast cancer is now the leading cause of death in the developed world for women aged 40-55. In the western world, more than 10 % of women will develop breast cancer at sometime during their lifetime, and this rate has increased dramatically in recent years. Increased levels of the aspartic proteinase cathepsin D were first reported in several human neoplastic tissues in the mid-eighties (Vetvicka *et al* 1999). These findings generated intense research to find a possible role for cathepsin D in neoplastic processes. These investigations demonstrated a strong predictive value for measurement of cathepsin D concentrations in breast cancer as well as many other types of tumour. Cathepsin D (E.C.3.4.23.5.) is a lysosomal aspartic proteinase unlike most of the other members of the aspartic proteinase family that are secretory proteins. Procathepsin D is sorted to the lysosomes due to the presence of a mannose-6-phosphate tag that is recognized by a mannose-6-phosphate receptor. Upon entering into the acidic lysosome, the single-chain procathepsin D (52 kDa) is activated to cathepsin D and subsequently to a mature two-chain cathepsin D (31 and 14 kDa, respectively). The two mannose-6-phosphate receptors involved in the lysosomal targeting of procathepsin D are both expressed intracellularly and on the outer cell membrane. The glycosylation is believed to be crucial for normal intracellular trafficking.

The fundamental role of cathepsin D is to degrade intracellular and internalized proteins. Another role suggested for Cathepsin D is in antigen processing and in enzymatic generation of peptide hormones. Cathepsin D is functional in a wide variety of tissues during their remodelling or regression, and in apoptosis. Within the mammary gland cathepsin D has been linked to the processing of the peptide hormone prolactin. The synthesis of cathepsin D is controlled by steroid

hormones, for example in breast cancer cell lines, cathepsin D expression is regulated by estrogens. Under normal physiological conditions both procathepsin D and cathepsin D are not secreted, they are found only intracellularly. In ER+ (estrogen receptor positive) cell lines, procathepsin D is secreted after estrogen stimulation and it is secreted constitutively in ER- cell lines. Increased levels of cathepsin D (both at the mRNA and protein levels) were first reported in several human neoplastic tissues in the mid-eighties (Fusek and Vetvicka 1994). There are indications that it may serve as a growth factor with the mitogenic region residing in the propart region of the zymogen. The design of antagonists to block the interaction of the activation peptide with its as yet unknown receptor maybe a valuable tool in breast cancer inhibition.

Malaria

The most virulent malaria parasite *Plasmodium falciparum* produces up to nine aspartic proteinases (the plasmepsins) and a closely related protein HAP. Three of these enzymes (plasmepsins I, II and IV) are located within the parasite food vacuole and are involved in digestion of host red cell haemoglobin by the parasite. Malarial parasites mature in human erythrocytes and, during occupation of these cells, the parasites degrade up to 80 % of the cell's haemoglobin. This provides essential nutrients for parasite growth and creates growing space for the parasite within the cell. Two aspartic proteinases, plasmepsins I and II from the malaria parasite *Plasmodium falciparum*, have already been identified as having a pivotal role in the degradation of haemoglobin, which occurs within the lysosomes of the parasite. Thus plasmepsins I and II have been identified as potential drug targets. HAP is also found in this organelle but its function is presently unknown. Inhibition of the plasmepsins can lead to parasite death providing a possible target for aspartic proteinase inhibitors. Cysteine and aspartic protease inhibitors are now under study as potential targets for anti-malarials. Lead compounds which inhibit either the parasites cysteine or aspartic proteinases have blocked *in vitro* parasite development at nanomolar concentrations and cured malaria-infected mice.

The Aspartic proteinase Endothiapepsin

Protein Structure

Endothiapepsin contains 330 amino acids in a single chain and has a molecular weight of 33.8 kDa. The full amino acid sequence was determined via chemical means by V. Barkholt (1987) and confirmed later by DNA sequencing (Razanamparany *et al.*, 1992; Choi *et al.*, 1993). Endothiapepsin contains a single disulphide bridge between cysteine residues 250 and 283. The amino acid sequence has signs of an internal repeat relating the two halves of the molecule. Their identity is greatest in the vicinity of the catalytic residues that occur in the two conserved Asp-Thr-Gly sequences. The secondary structure of endothiapepsin possesses the same fold as other pepsin like aspartic proteinases being largely β-sheet with small areas of α helix and consisting of two topologically related lobes of approximately 170 amino acids each.

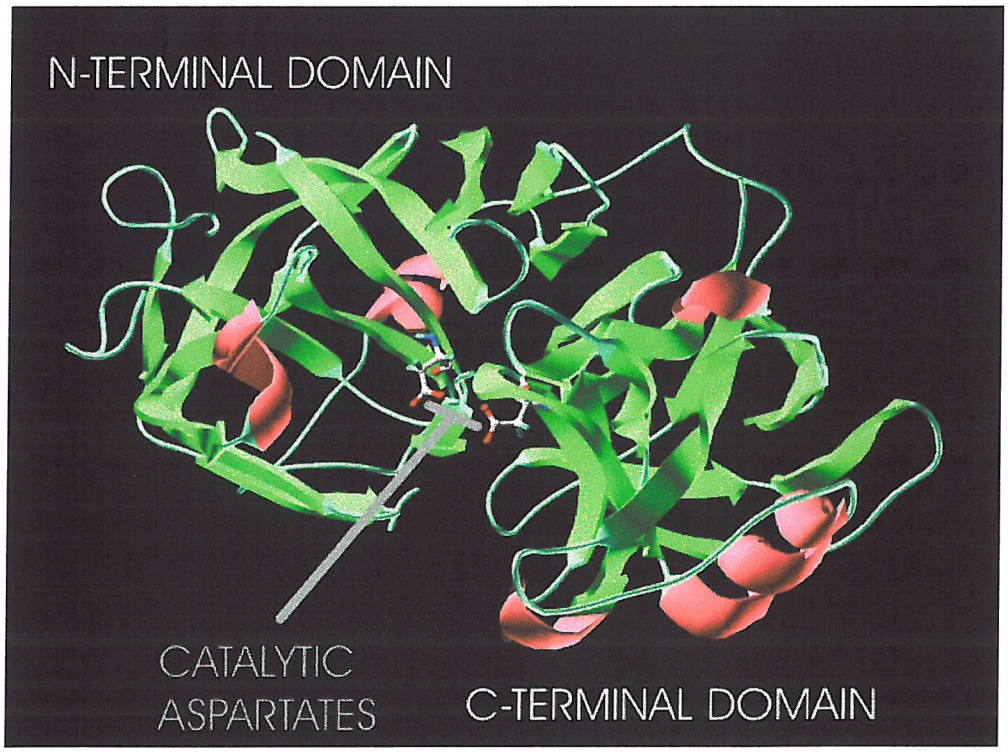


Figure 1.00 Showing the secondary protein structure of endothiapepsin, the catalytic aspartates are shown in ball and stick with β sheets in green and α helices in pink.

Most of the structures of the retroviral and pepsin like proteinses are composed of β-sheets arranged in a simple, symmetrical way. Each lobe comprises of two similar motifs formed from anti-parallel β strands a, b ,c and d for the first lobe and a', b', c' and d' for the second (Dunn 1991).

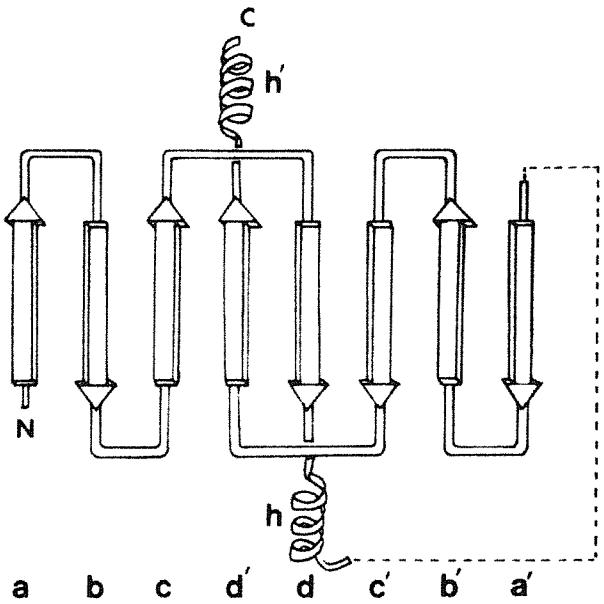


Figure 1.01 A schematic diagram of the strands and helices of the super-secondary structure that is common between retroviral proteinases and pepsins. Diagram taken from Dunn 1991.

These are organised together in a disordered sheet. Strands c and d' and strands c' and d form two pairs of parallel strands. Strands b and c and strands b' and c' form anti-parallel β-hairpins that are folded over sheet 1 and hydrogen bonded together around an intra-domain two fold axis to give a second sheet (sheet 2), which is orthogonal to the first. Much of strand a of the first motif is displaced from the main sheet and forms an anti-parallel β-sheet with the carboxy-terminal strands of the subunit or lobe, and their equivalents in the second lobe (sheet 3). Two carboxyl-terminal strands of each lobe contribute to a six stranded anti-parallel β-sheet. The Ψ structure which contains the active site is formed from the c d' and d strands.

The active site resides in a pronounced cleft between the two lobes of the protein. The base of this cleft is made of β-strands forming two abutting Ψ structures, which contain the catalytic aspartate residues (32 and 215 in endothiapepsin). The carboxyl groups of the catalytic aspartates are held coplanar and within hydrogen bonding distance by an intricate arrangement of hydrogen bonds involving the surrounding main chain and conserved amino acid side chain groups. Kinetic studies on *Rhizopus* pepsin in which Ser 35 and Thr 218 were mutated to alanines have shown that these residues play an important role in catalysis (Lin *et al* 1992), however it is not fully understood how the these residues enhance the catalytic rate. Hartree fock calculations carried out by Beveridge 1998 indicate that Ser 35 lowers the pK_a of Asp 32 facilitating the formation of the *gem*-diol intermediate. The pH vs activity profiles for endothiapepsin and other aspartic proteinases are bell shaped (Fruton 1976) indicating the active site carries a formal charge of −1 to effect hydrolysis.

A solvent molecule bound tightly to both carboxyls by hydrogen bonds is found in all native aspartic proteinase crystal structures. The aspartyl carboxyl groups and the water molecule all lie in a single plane. The bound water molecule is within hydrogen bonding distance of all four carboxyl oxygen atoms. Synthetic and naturally occurring inhibitors based on transition state analogues have been studied by X-ray crystallography and are shown to bind to the active site cleft in an extended conformation with up to 10 residues of the inhibitor interacting with the active site cleft (Bailey *et al.*, 1993; Blundell *et al.*, 1987; Cooper *et al.*, 1987; 1989; 1992; Foundling *et al.*, 1987; Lunney *et al.*, 1993; Sali *et al.*, 1989; 1990; Veerapandian *et al.*, 1990; 1992). The hydrogen bond acceptors and donors which help position the substrate's main chain in the active site cleft are largely conserved from enzyme to enzyme. This implies that the largest determinants of specificity are the van der Waals contacts between the enzyme and the ligand's side chains. The numbering system for substrates uses the dipeptide in the substrate that is cleaved as its starting point. The first residue on the N terminal fragment after the cleavage site is called P_1 while the first residue on the C terminal end of the substrate after the cleaved residue is called $P_{1'}$. The central region of an

inhibitor except for P_2 is almost completely shielded from the solvent by hydrophobic binding pockets in the protein and an anti-parallel β-hairpin formed by residues 71-82. This structure is known as the flap and it shields the active site from the solvent. The hydroxyl groups present in transition state analogues bind by hydrogen bonds to the catalytic aspartates and occupy the same position as the water molecule in the uncomplexed enzyme. This water molecule has been implicated in catalysis by Suguna *et al.*, (1987) who suggested that it becomes partly displaced when the substrate binds and is polarised by one of the aspartate carboxyls. The water then nucleophilically attacks the scissile bond carbonyl group. This mechanism has been refined based on the high resolution structures of geminal diol transition state analogues (difluoroketone: $-C(OH)_2-CF_2-$) in which both hydroxyls of the putative transition state are represented (Veerapandian *et al.*, 1992; James *et al.*, 1992). A detailed comparison of the X-ray structures of 21 different inhibitor complexes is given by Bailey and Cooper (1994). The strongly conserved binding interactions at P_4, P_3, P_1, $P_{1'}$ and $P_{2'}$ compared to the weaker binding and unfavourable geometry of the P_2 residue may indicate that the enzyme possibly strains bound substrates at P_1 and P_2 towards the geometry found in complexes with transition state isoteres. Intriguing effects are also seen wherein the side chains of the inhibitors can adopt different conformations to compensate for greater or lesser occupation of the neighbouring subsites in different complexes.

Enzyme Mechanism

One proposed mechanism for endothiapepsin is based on a study of the enzyme's structure when co-crystallised with an inhibitor containing a *gem*-diol unit. This unit contains both of the hydroxyls thought to exist in the tetrahedral intermediate. Endothiapepsin was co-crystallised with a *gem*-diol containing inhibitor (CP-81,282) and after refinement of the structure to 2.0 Å resolution an R_{factor} of 0.18 was obtained (Veerapandian *et al* 1992). This structure led to the model of the active site shown in Figure 1.02. The proton positions in Figure 1.02 are based on

the positions of the heavier polar atoms. An X-ray structure at 2.0 Å is not at a high enough resolution to determine the positions of protons directly.

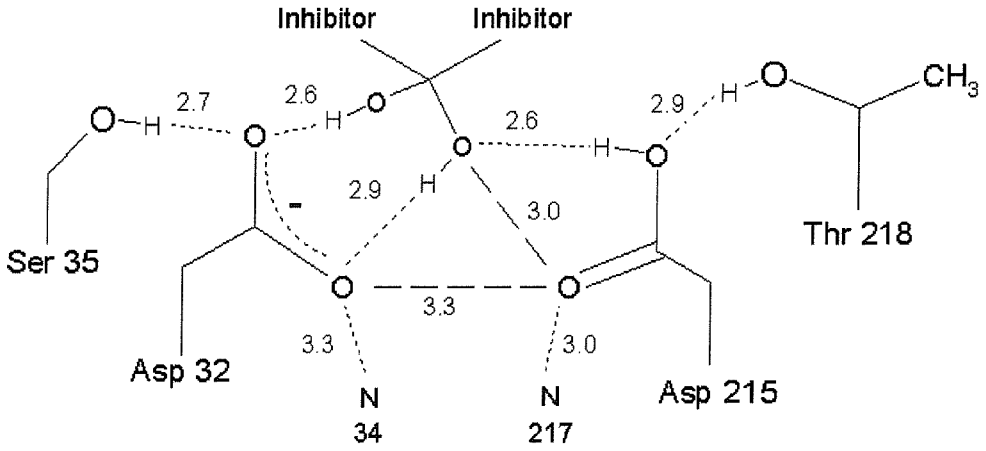


Figure 1.02 Model of the inhibitor CP-81,282 binding to the active site of endothiapepsin, dotted lines indicate hydrogen bonds with distances given in Ångstroms. Diagram adapted from Veerapandian *et al* 1992.

The inner carboxyl oxygen atoms of both Asp 32 and 215 are hydrogen-bonded to the peptide nitrogen atoms of residues glycine 34 and glycine 217. The outer carboxyl oxygens of Asp 32 and Asp 215 make weaker hydrogen bonds with the side chain hydroxyl groups of serine 35 and threonine 218 respectively. The side chains of the threonines at the active site are directed towards the hydrophobic core of the protein where they are involved in a "firemans grip" hydrogen bonding pattern which aids the structural stability of the catalytic centre (Blundell *et al* 1990). These results were used to form a proposed mechanism for endothiapepsin, which is shown in Figure 1.03.

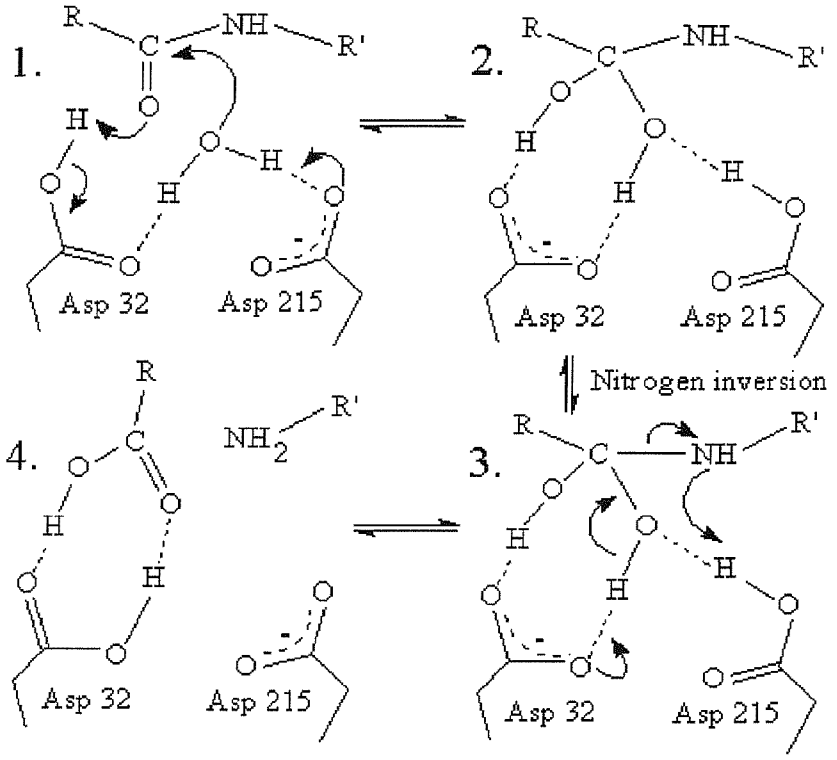


Figure 1.03 Mechanism of peptide bond cleavage by endothiapepsin as proposed by Veerapandian *et al* 1992.

In the first stage of the proposed reaction mechanism (stage 1 to 2) the scissile bond carbonyl is protonated by Asp 32 then attacked by a water molecule polarised into a nucleophilic state by Asp 215. This forms a tetrahedral intermediate (stage 2). In stage 3 the second tetrahedral intermediate is formed by either a nitrogen inversion or an 60° rotation about the $C(OH)_2$-N bond. The final reaction to produce the products (Stage 4) involves the protonation of the amide by Asp 215. This destabilises the tetrahedral intermediate by the removal of a hydrogen bond from Asp 215 to the statine like hydroxyl. This action breaks down the tetrahedral intermediate severing the peptide bond.

At the start of this project there where 29 endothiapepsin structures in the protein database all of which have been solved using X-ray diffraction to a maximum resolution of 1.6 Å and refined isotropically. The catalytic mechanism of the enzyme was still uncertain as there was no direct experimental evidence for any hydrogen positions in the active site with or without an inhibitor bound. The aim of this project was to utilise the unique properties of neutron scattering to locate the hydrogen atoms in the active site when a transition state analogue inhibitor (H261) is bound at the active site. Another aim of the project was to collect X-ray diffraction data to atomic resolution <1.2 Å of endothiapepsin bound to a transition state analogue inhibitor to enable the anisotropic modelling of the protein. At this resolution the protonation states of the aspartate groups in the active site could be determined from differences in the carboxyl bond lengths. At this resolution there could also be some weak electron density for the hydrogen atoms themselves. The combination of these techniques should enable the determination of the active site hydrogens thus eluding by which of the proposed catalytic mechanisms the enzyme operates.

Properties of Hydrogen bonds

As can be seen in Figure 1.02 some of the hydrogen bond lengths (donor-acceptor distances) in the active site are ~2.6 Å. Although the resolution of these structures is not at high enough resolution to give convincing bond length estimates they do suggest the possibility of short hydrogen bonds in the active site. Hydrogen bonds as short as 2.5 to 2.6 Å are referred to as low barrier hydrogen bonds (LBHB) since the proximity of the donor and acceptor reduces the energy barrier which normally prevents transfer of the hydrogen atom from the donor to the acceptor group (Cleland *et al* 1998). Thus rapid exchange of the proton between the donor and the acceptor atoms is facilitated, this has been proposed as being an important effect in a number of other enzymes including citrate synthase and serine proteinases. The strength of a hydrogen bond depends on its length and linearity and the nature of its microenvironment. In water, the hydrogen-bonded oxygens are separated by a distance of around 2.8 Å with the ΔH of formation being ~5

kcal mol^{-1}. The hydrogen bonds in water are weak however because of the poor p*K* match between the participating oxygen atoms. Because the p*K*s of H_3O^+ and H_2O are −1.7 and 15.7, respectively, the proton in the structure $H_2O\cdots$H-OH is tightly associated with the OH^- group as a water molecule (Cleland *et al* 1998). In the gas phase, where the dielectric constant is low, hydrogen bonds between heteroatoms with matched p*K*s can be very short and strong, and experimental as well as calculated values of ΔH of formation can approach 25 or 30 kcal mol^{-1}. In organic solvents, strong hydrogen bonds can also form although the ΔH of formation probably never exceeds 20 kcal mol^{-1}. Because the active site of an enzyme is no longer aqueous once it has closed around a substrate, as is the case in endothiapepsin the properties of hydrogen bonds in organic solvents are highly pertinent to enzymatic catalysis. What happens energetically as hydrogen bonds become shortened can be seen in Figure 1.04. Structure *A* represents the situation in water in which the hydrogen is firmly attached to either the left-hand or right-hand oxygen and is more loosely bonded to the other one, with an O-O distance of around 2.8 Å. There is an energy barrier between the two possible positions of the hydrogen, as shown in Figure 1.04. Such a hydrogen bond is essentially electrostatic, and the covalent O-H bond is the usual 0.9-1.0 Å in length. As the overall O-O distance is shortened, the energy barrier drops until it reaches the zero point energy level at an O-O distance of ~2.5 Å (Fig. 1.04 *B*); this is a LBHB. The ΔH of formation has increased to around 15-20 kcal/mol, and the hydrogen can now move freely between the two oxygens. In crystals containing LBHBs, neutron diffraction shows the hydrogen atom diffusely distributed with its average position in the centre (Steiner and Saenger 1994). LBHBs are largely covalent, and there are a number of possible structures between *A* and *B* in Fig. 1.04; however, the covalent O-H bond becomes longer and the overall covalent character of the hydrogen bond increases as the hydrogen bond becomes shorter and stronger (Steiner and Saenger 1994). Further shortening leads to the limit of 2.29 Å and structure *C* a single well hydrogen bond shown in Fig. 1.04.

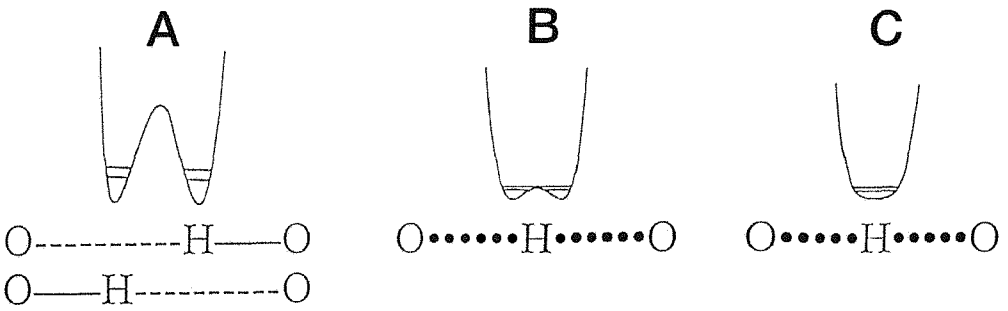


Figure 1.04 Energy diagrams for hydrogen bonds between groups of equal pK_a . **A**, weak hydrogen bond with O-O distance of 2.8 Å; the two positions of the hydrogen are shown. **B**, low barrier hydrogen bond of length 2.55 Å; the hydrogen is diffusely distributed, with average position in the centre. **C**, single-well hydrogen bond with length 2.29 Å. The *upper* and *lower horizontal lines* are zero point energy levels for hydrogen and deuterium, and the *curves* define the energetic barrier to changes in bond length. The distances are to scale (taken from Cleland *et al* 1998).

It is also possible to detect low barrier hydrogen bonds in protein inhibitor complexes using the presence of signals in the low field proton NMR signals (Cassidy *et al* 1997).

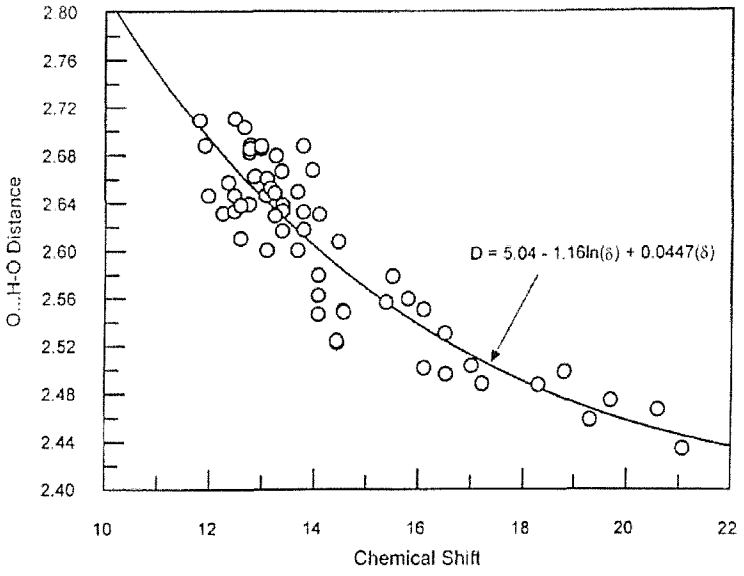


Figure 1.05 showing the correlation of O-H•••O hydrogen bond distances with 1H chemical shifts from solid state NMR crystalline amino acids. (Taken from Mcdermott and Ridenour 1996)

A proton involved in a LBHB has a NMR chemical shift far downfield, typically between 16-21 ppm. It can be observed in aqueous solution by application of appropriate water suppression pulse sequences when the exchange rate is slower than the spectrometer frequency (Cleland *et al* 1998). There are two main properties that distinguish low barrier hydrogen bonds from normal hydrogen bonds. The first is a short length with the second being the exclusion of local solvation effects. There is evidence that LBHB formation does not require as close a match of proton affinity (pK_a) as was previously thought (Shan *et al* 1996).

Low barrier hydrogen bonds in enzyme catalysis

There are a few enzymes in which LBHBs have been identified within the active site of the enzyme such as the serine proteinases and Δ^5-3-ketosteroid Isomerases (KSI). The role of LBHBs in the catalytic mechanism of KSIs is well characterised and has been recently reviewed by Ha *et al.*, 2001. KSI has been intensively studied for the last 45 years as a prototype for understanding the chemical and thermodynamical aspects of enzyme catalysed C-H bond cleavage. The highly apolar active site consists of three catalytic residues, two of which (Asp 38 and Tyr 14) were identified by mutagenesis studies. The third residue Asp 99 was identified due to its location and hydrogen bonding contacts within the active site. In the proposed mechanism the LBHB between Tyr 14 and the ketosteroid helps to stabilise a negative charge on the tetrahedral intermediate.

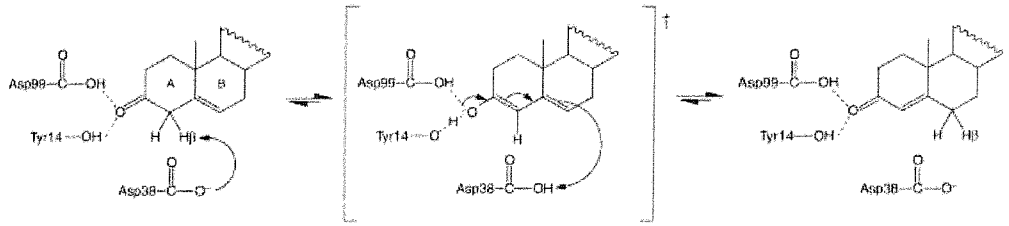


Figure 1.06 The proposed mechanism of KSI. The reaction is initiated by the abstraction of a proton from the steroid substrate by Asp 38. The resulting dienolate intermediate is stabilized by two hydrogen bonds provided by Tyr 14 and Asp 99. Subsequent reketonization results in protonation at the C6 position by Asp 38. Figure taken from Han *et al* 2001.

While there is good NMR and biochemical evidence to support the presence of a LBHB in the active site, the resolution of the crystal structures at around 2.0 Å do not give accurate enough bond lengths to allow definite identification of LBHBs. Another problem with the mechanism outlined in Figure 1.06 is the energy barrier (~11kcal/mol) for transfer of the proton from the steroid to Asp 38; how this energy barrier is overcome represents a major mechanistic question. One possibility that has been suggested is the active site alters the pK_a of the proton donor and acceptor groups to reduce the energy barrier required for proton transfer.

The role of LBHBs in enzyme catalysis is hotly debated. Recent results from studies of serine proteases prompt caution in interpreting the role of LBHBs in catalysis. Mutation of the aspartate in subtilisin that forms a LBHB bond to the histidine in the catalytic triad yields variable results depending on the nature of the replacement. An Asp⟶Asn mutation leads to more than a 10000-fold loss in catalytic activity (k_{cat}/K_m), whereas an Asp⟶Cys mutation causes less than a 50-fold decrease, despite the absence of a low-barrier hydrogen bond in the mutant (Stratton *et al* 2001).

Background to neutron protein crystallography studies

The number of protein structures solved by neutron diffraction is less than twenty at the time of writing. there are a number of reasons for this but the primary restriction is caused by the large crystals that where required for neutron diffraction >3mm^3. Most proteins will not form crystals of this size prohibiting their use in neutron diffraction experiments. There are also a number of other limitations that affect a crystal's suitability for neutron diffraction; these are discussed in detail in chapter 3. Another factor limiting the number of neutron diffraction studies is the extended beam time required to collect a complete data set caused by the low flux present in monochromatic neutron beams. However recent developments in the field of neutron diffraction including the use of polychromatic neutron beams and improved detector arrays have reduced the size

of crystal required to 1mm^3 increasing the feasibility of neutron diffraction studies. Recent proteins studied by neutron diffraction include lysozyme (Bon *et al* 1999) and concanavalin A (Habash *et al* 2000). Both these studies centred on identification and geometry of water molecules bound to the protein surface. Earlier monochromatic neutron diffraction studies by Kossiakoff and Spencer (1980) allowed the direct determination of the protonation states of Asp 102 and His 57 of the catalytic triad in the tetrahedral intermediate of trypsin Kossiakoff and Spencer (1981). The neutron work undertaken here is comparable to the work of Kossiakoff and Spencer in that it is the catalytic mechanism of the enzyme that is the primary interest. In addition, this thesis presents the largest protein to be studied at high resolution using neutron diffraction. Previous studies on the protonation states of aspartates and glutamates have been conducted by Deacon *et al* 1997 on concanavalin A using 0.94 Å X-ray data.

Chapter 2

Protein purification, crystallisation and transition state analogue inhibitors

Protein purification

For a protein crystal to form a high level of homogeneity is usually required. Thus high levels of protein purity with sequence integrity need to be achieved. There are a large number of techniques that are employed to purify proteins. The techniques and protocols used to purify the aspartic proteinase endothiapepsin are similar to that of other fungal aspartic proteinases. The initial stages of purification consist of a salt fractionation using ammonium sulphate as the precipitant followed by column chromatography techniques such as gel filtration and ion exchange. Desalting of the protein sample is carried out by dialysis with concentration of the protein being performed by ultra filtration. All stages in the extraction and purification of endothiapepsin are carried out at 4°C to minimise the heterogeneity caused by the autolysis of the protein.

The starting point in the purification of endothiapepsin is fungal rennet also called "Sure-Curd" which is dissolved in 0.1M sodium acetate buffer at pH 5.6. The material is then dialysed and twice fractionated using a 60 % saturated ammonium sulphate solution. Precipitation with ammonium sulphate or polyethylene glycols is often used as the starting point in protein purification because it can be used to crudely separate proteins from nucleic acids. Nucleic acids are highly charged polyanions that can compromise the efficiency of column chromatography techniques e.g. they can saturate ion exchange columns. The pellet resulting from precipitation is then dissolved in 0.1M sodium acetate buffer at pH 5.6 as before and dialysed again before being run down a G-100 sephadex column which fractionates proteins by size. Fractions containing the enzyme are collected from the column and pooled, concentrated and dialysed ready for the next stage in the purification. This consists of ion exchange chromatography in a sephadex A-50 column. The protein solution is loaded onto the sephadex A-50 column and elution takes place by a linear increase in the concentration of sodium acetate at pH 4.6 from 0.05 to 0.5M. The A-50 ion exchange column separates proteins via their ability to form ionic interactions with the positively charged column material. The proteinase solution is applied at low ionic strength at a suitable pH (around

5.5) at which ionic interactions between the column and the protein occur. The bound proteins are then washed off the column by increasing the amount of anions in the eluting solution that compete for the bindings sites on the column with the proteins in the column. The use of a linearly increasing gradient of anions thus permits proteins to be separated on their ability to bind to the ion exchange column. Pooled fractions from the column are again dialysed and then freeze dried. The freeze dried protein is then redissolved in 0.05M formate buffer at pH 4.6 and then run down a G-150 molecular sieve column. The G150 gel filtration columns fractionate globular proteins with molecular weights ranging between 4 to 150 kDa. This separation is based on the protein size, small proteins are able to enter and pass through the beads in the gel and thus their passage through the gel is retarded compared to larger proteins which are unable to enter the beads and therefore elute from the column first.

The enzyme obtained is then assayed by use of a milk clotting assay. In this assay 0.6g of dried skimmed milk is mixed with 5 ml's of distilled water to which is added 225 ml of 0.2M sodium acetate buffer at pH 5.3 and 20 ml of 10mM $CaCl_2$ to give a final volume of 250 ml of stock solution. The assay is performed at room temperature by mixing 3 ml of stock solution with 5-10 µl of the eluted enzyme fraction. This solution is placed in a visible range cuvette and the change in absorbance at 500 nm is measured over the 0 to 2 optical density units. The change in absorbance which indicates proteinase enzyme activity follows a sigmoidal curve and the clotting time is measured as the time between the addition of the enzyme and the inflection point of the absorbance curve. The clotting time is inversely proportional to the proteinase concentration thus a standard curve of clotting time using a known enzyme concentration allows estimation of the proteinase concentration to be calculated to determine the yield and purity of the enzyme.

Finally to check the purity of the final protein around 3 µl of the final protein solution are run in a polyacrylamide gel-electrophoresis experiment. In this type of experiment the protein sample is boilied in a detergent such as sodium dodecyl sulphate (SDS) prior to loading onto the gel. This denatures all the proteins within the sample and causes them to assume a rod like shape. Most proteins bind SDS in the same ratio of 1.4g of SDS per 1g of protein. Furthermore since SDS imparts a negative charge when it binds to proteins it shields the intrinsic charge of the protein so that SDS treated proteins have identical charge to mass ratios and similar shapes. Hence proteins can be gel filtered on a polyacrylamide gel when a positive current is used to move the negatively charged protein down the gel achieving separation by molecular weight. After staining with an agent known to bind to proteins such as comassie blue, a number of bands are visible on the gel and these correspond with the molecular weight of protein(s) in the sample. The weight of the proteins in these bands can be estimated by comparison to lanes of the gel that contain marker fragments of known molecular weight. A single band at 35 kDa would indicate a fairly pure sample of endothiapepsin. A high level of purity is required for crystallisation as it is easier to achieve the high protein concentrations < 10 mg/ml usually needed for crystallisation, the behaviour of highly purified protein is also more reproducible. Degradation of the protein during storage will also be minimised if all unwanted proteinases have been removed.

Protein crystallisation

Endothiapepsin was co-crystallised with each transition state inhibitor using ammonium sulphate as a precipitant at pH 4.5 via an adaptation of the original batch method of Moews and Bunn 1970. This involved the addition of a 3 fold molar excess of inhibitor to enzyme at a concentration of 2.0 mg/ml in 100 mM sodium acetate buffer at pH 4.5. Finely ground ammonium sulphate was added until slight turbidity was apparent (around 55 % saturation) at which point the solution was micropore filtered and, if necessary a few drops of acetone were added to clear any remaining turbidity. 2ml of this solution were placed in small

glass vials in which crystals were grown. In the batch method concentrated protein is mixed with concentrated precipitant to produce a final concentration that is supersaturated in terms of the solute protein and therefore leads to crystallisation. The batch method typically provides larger crystals due to larger volumes of protein present and the reduced chance of impurities diffusing to the face of the crystal halting growth. This technique consumes large amounts of protein and thus is not generally used to screen initial conditions for protein crystallisation. The PD-135,040 co-crystals were grown via the hanging drop vapour diffusion method in which 5 to 20 µl of a 2 mg/ml concentration of protein solution at pH 5.0 in a 0.1M sodium acetate buffer were placed on silconised cover slips. Each cover slip was then sealed with high vacuum grease over a well containing a saturated ammonium sulphate solution at a range of different pH values 4.3, 4.5 and 4.7 all in a 0.1M sodium acetate buffer. Small crystals (0.1 x 0.3 x 0.3mm) appeared after three weeks.

Endothiapepsin crystals grow in two different crystalline forms. One of the forms has a high solvent content (55 %), is more prone to macroscopic twinning and diffracts more weakly than the other form (Badasso *et al* 1992). The co-crystals used in this work belong to the better ordered second crystal form (Cooper and Myles 2000) which have a solvent content of 39 % and are less prone to macroscopic twinning except for the PD-135,040 crystals which are in the higher solvent content unit cell. At the beginning of the project a number of endothiapepsin inhibitor co-complex crystals were available for data collection, these had been grown some years ago by Dr. Jon Cooper. The endothiapepsin PD-135,040 were freshly grown during the PhD.

Endothiapepsin H261 complex

The crystals of endothiapepsin used in the neutron work were complexed with inhibitor H261 via an adaptation of the original method of Moews and Bunn (1970) detailed earlier. Crystal formation took several weeks with some crystals being stored to grow in the mother liquor for a period of eleven years; the largest

of these crystals have the dimensions of 1.8 x 1.4 x 1.4 mm (Figure 2.04). In order to reduce the contribution of the large incoherent neutron scattering cross section of hydrogen to the experimental background, crystals of the endothiapepsin-H261 complex were subjected to hydrogen-deuterium (H-D) exchange by vapour diffusion for several months prior to data collection. The H-D exchange involved mounting selected crystals in large diameter capillaries with deuterated mother liquor which were then sealed with dental wax. The crystal was equilibrated against successive changes of a 90 % deuterated mother liquor solution within the capillary and, finally, against several changes of a 95 % deuterated solution. The exchange of deuterated mother liquor was done by the melting of the one of the dental seals on the capillary and drawing out the deutertated mother liquor using a syringe, fresh deuterated mother liquor was the added and the capillary resealed. This method has been shown to be effective in other studies (Bon *et al* 1999). The decision to use vapour diffusion rather than soaking directly in D_2O was to avoid shocking the crystals since only a few were large enough for neutron data collection and they were 11 years old at the time of the experiment. The techniques used for mounting crystals in capillaries are shown in Figure 2.00.

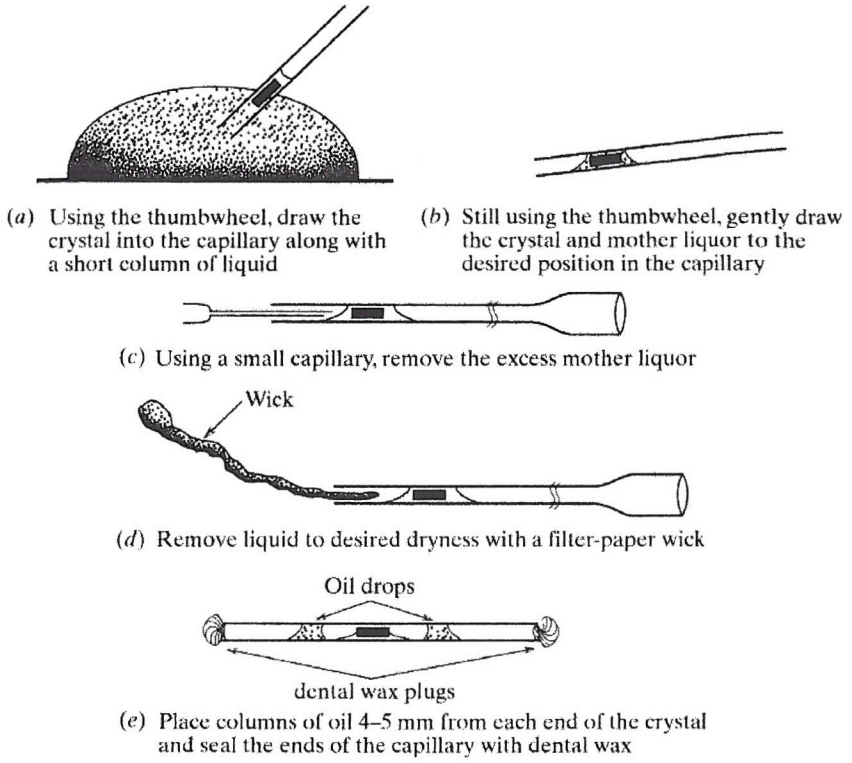


Figure 2.00 Illustrating the techniques used to mount a protein crystal in a capillary. Diagram taken from Carrell and Glusker 2001.

Data was collected using the LADI detector at ILL in Grenoble with the crystal at room temperature over a total period of three months. H261 is very potent inhibitor of endothiapepsin, it has an inhibition constant (K_i) of 0.7 nM (Veerapandian *et al* 1990) and is composed of eight residues and has the sequence Boc-His-Pro-Phe-His-LOV-Ile-His (Figure 2.01).

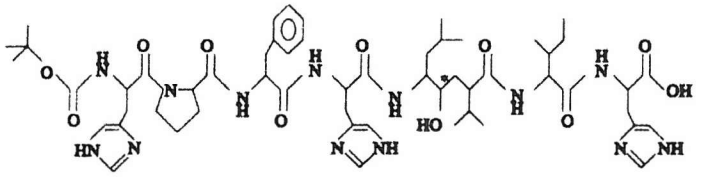


Figure 2.01 showing the chemical composition of H261, * indicates the inhibitory group.

The LOV group (leucine hydroxyethylene valine analogue) is a non-hydrolysable analogue of the natural substrate and possesses an isotere of the transition state, while the BOC group is a tertiary butyl oxycarbonyl group. The -CHOH-CH_2- of LOV mimics one of the hydroxyls in the proposed geminal-diol intermediate and the -CH_2- of LOV mimics its tetrahedral nitrogen. It is thought that the interactions the inhibitor makes with the protein will be similar to those formed transiently in the tetrahedral intermediate state during catalysis. The single hydroxyl group present in the LOV occupies the position of the bound water in the inactive enzyme; the bound water itself is displaced when LOV binds.

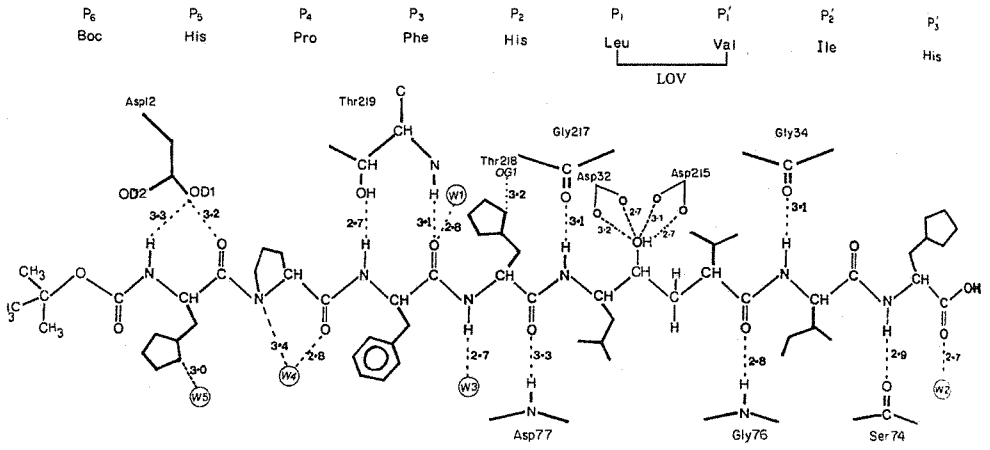


Figure 2.02 showing the possible hydrogen bonds between H261 and endothiapepsin, donor acceptor distances shorter than 3.4 Å are shown with broken lines (taken from Veerapandian *et al* 1990).

The H261 inhibitor forms an extended β sheet conformation in the active site and forms hydrogen bonds and van der Waals interactions from its main chain to residues on each lobe of the enzyme (Figure 2.03). The central region of the inhibitor also interacts with a region of the protein called the flap; this interaction helps to bind the substrate as tightly as possible. The flap is made up from residues 71-82, which form an anti-parallel β sheet. This part of the protein is mobile and changes conformation when the substrate binds to the enzyme covering the active site shielding it from the solvent.

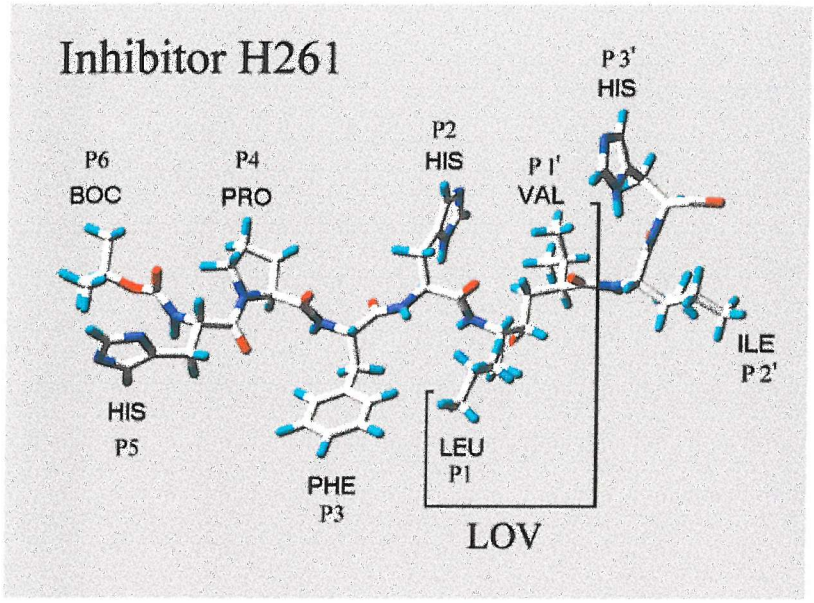


Figure 2.03 showing the 3D structure of H261 when bound in the active site of endothiapepsin. The highlighted LOV group occupies both the P1 and P1' binding sites.

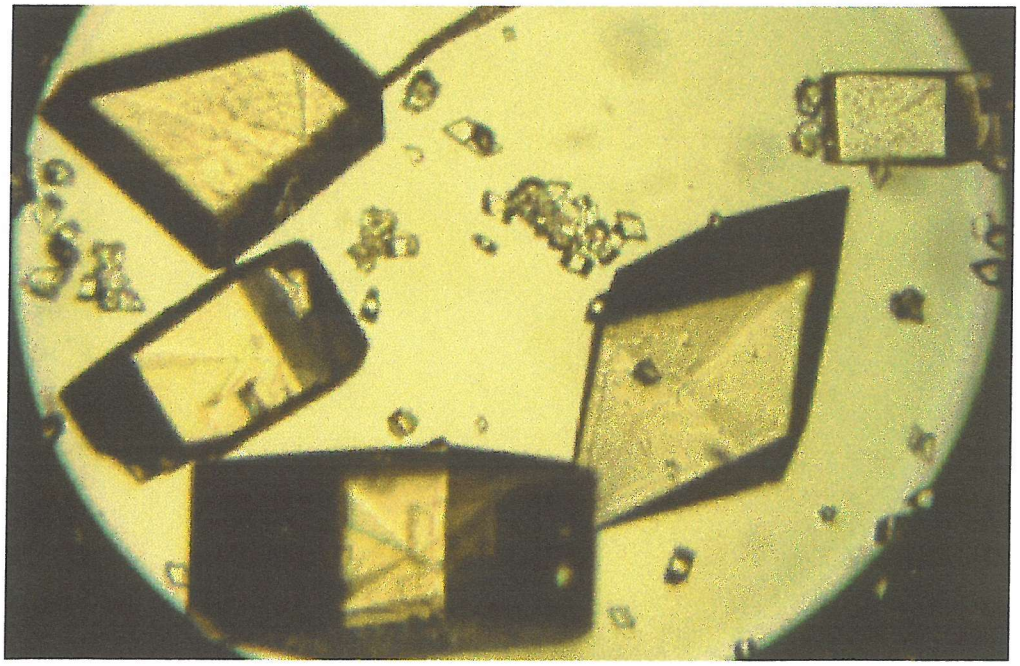


Figure 2.04 Showing endothiapepsin protein crystals co-crystallised with H261. The smaller crystals are suitable for X-ray diffraction while the larger crystals with a volume of around $3mm^3$ are suitable for neutron diffraction.

Endothiapepsin H189, H256, CP-80,794, PD-129,541,PD-130,328 and PD-135,040 complexes

High resolution X-ray data was also collected on six endothiapepsin inhibitor co-crystals (Figures 2.06, 2.08, 2.10, 2.12, 2.14 and 2.16). These crystals were again grown via an adaptation of the original method of Moews and Bunn but had smaller dimensions than the crystals used in neutron diffraction. These crystals where flash cooled using 40 % glycerol as a cyroprotectant and cryo-cooled with liquid nitrogen during storage and data collection and had the same non-native unit cell as the H261 crystal.

The phosphinic acid based inhibitor code name PD-130,328 (Figure 2.05) has the sequence Boc-Phe-His-PST-DCI with the inhibitory residue being a phosphostatine [-P(O)OH-CH_2-] (PST). The DCI residue (des-carboxy-Isoleucine) in the inhibitor makes extensive interactions with residues in the "flap" of the protein.

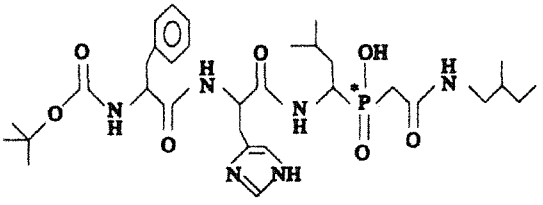


Figure 2.05 Showing the chemical composition of the PD-130,328 inhibitor, * indicates the inhibitory group.

The structure of this co-crystal had already been determined at 1.9 Å (C. Dealwis 1993 PhD Thesis), however it was though that it would be useful to collect a data set at higher resolution. The inhibitor itself has a relatively high K_i of 110 nM. This is thought to be because the phosphostatine may be less well accommodated in the active site than a hydroxyl group (Lunney *et al* 1993). Indeed a shift of Asp 215 relative to other inhibitor complexes has been seen by Lunney *et al* (1993), the energy cost in disrupting this region may be reflected in the weaker binding.

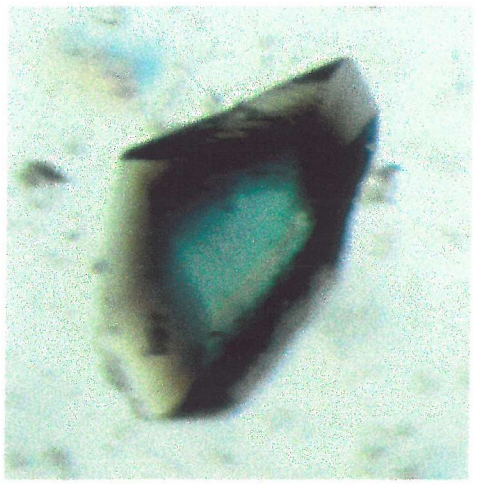


Figure 2.06 Showing a crystal of endothiapepsin co-crystallised with PD-130,328. The crystals produced with this inhibitor are fairly large at around $1 mm^3$.

H189 [Pro-His-Pro-Phe-His-Sta-(statyl)-Val-Ile-His-Lys] shown in Figure 2.07 is an excellent endothiapepsin inhibitor with a K_i of 1 nM (*Bailey et al* 1993), the inhibitor itself is an analogue of human angiotensinogen. The inhibitory statine (Sta) residue has been shown to replace both the P_1 and $P_{1'}$ residues (Bailey *et al* 1993). Its hydroxyl group forms hydrogen bonds with both of the catalytic aspartates.

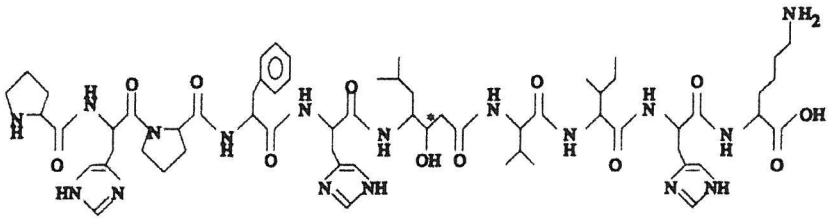


Figure 2.07 Showing the chemical composition of H189, * indicates inhibitory group.

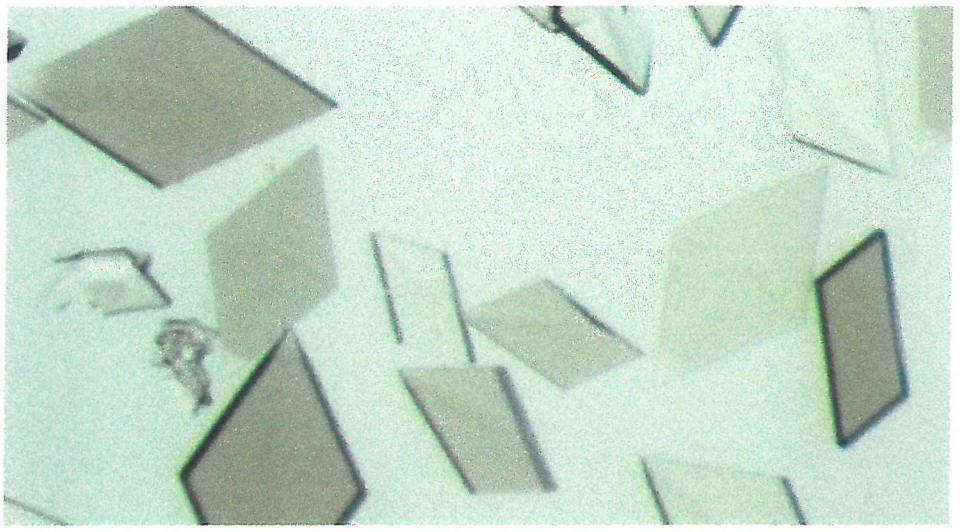


Figure 2.08 Showing the plate-like crystals of endothiapepsin co-crystallised with H189.

CP-80,794 [-Morpholino-Phe-Met-Norstatine-] (Figure 2.09) is a statine like analogue inhibitor which is shorter than H189. The tetrahedral intermediate is mimicked by the norstatine residue that contains a single inhibitory hydroxyl which binds to the active site aspartates.

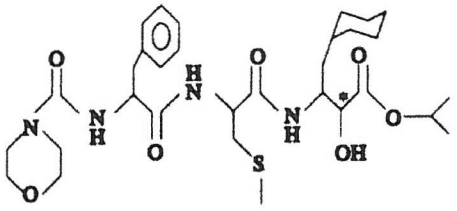


Figure 2.09 Showing the chemical composition of the inhibitor CP-80,794, * indicates inhibitory group.

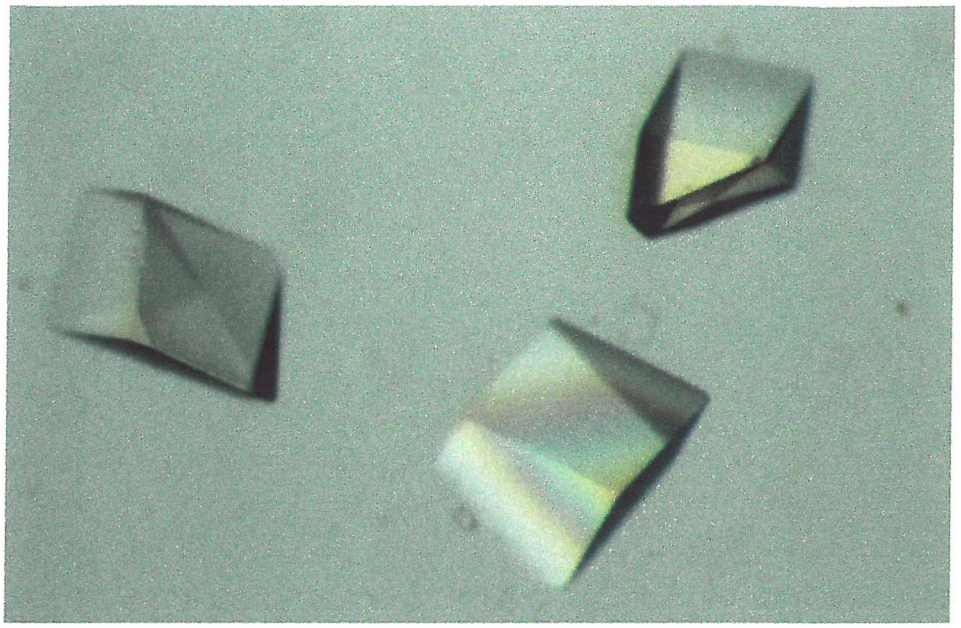


Figure 2.10 Showing crystals of endothiapepsin co-crystallised with CP-80,794.

PD-129,541 (Figure 2.11) is a cyclic peptide inhibitor which contains an inhibitory statine residue it contains a bis-[(1-naphthyl)methyl]acetic acid group (BNMA) which spans the P_4 and P_3 residues. The structure of this inhibitor bound to endothiapepsin had not been solved previously, however a structure of this inhibitor bound to Saccharopepsin (a yeast aspartic proteinase similar to endothiapepsin) has been solved to a resolution of 2.5 Å (Cronin *et al* 2000).

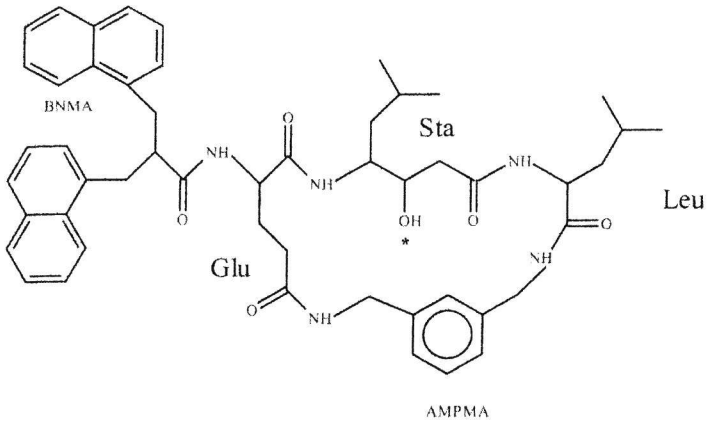


Figure 2.11 Showing the chemical composition of the inhibitor PD-129,541, * indicates the inhibitory group.

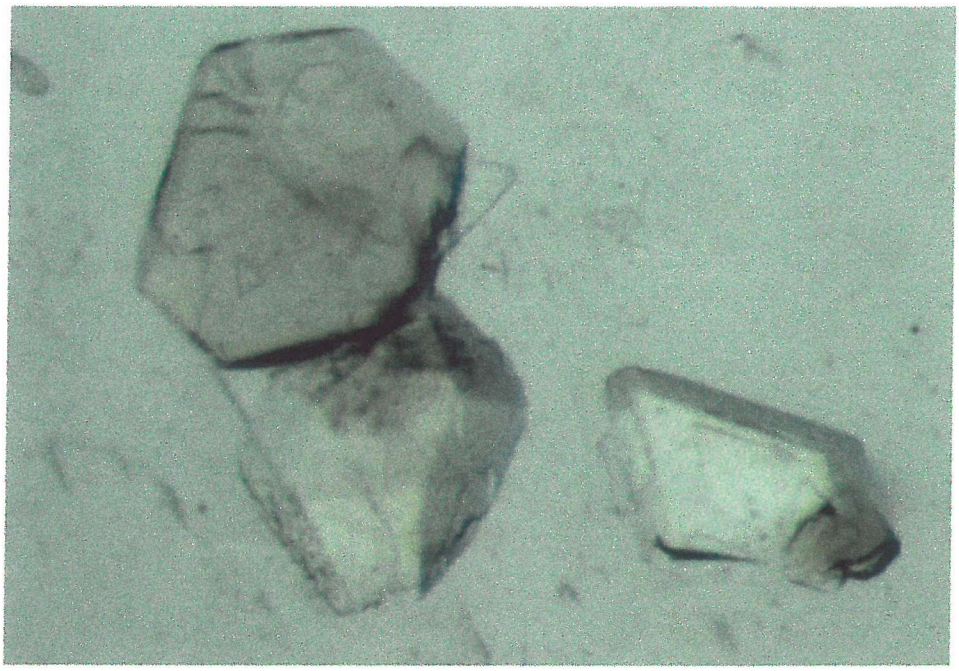


Figure 2.12 Showing crystals of endothiapepsin co-crystallised with PD-129,541.

H256 (Figure 2.13) possesses a reduced bond analogue in which the scissile peptide bond is replaced by a $-CH_2-NH-$ group. The K_i of this inhibitor with endothiapepsin is 60 nM somewhat higher than that of the statine based inhibitor H189 but somewhat lower than that of PD-130,328.

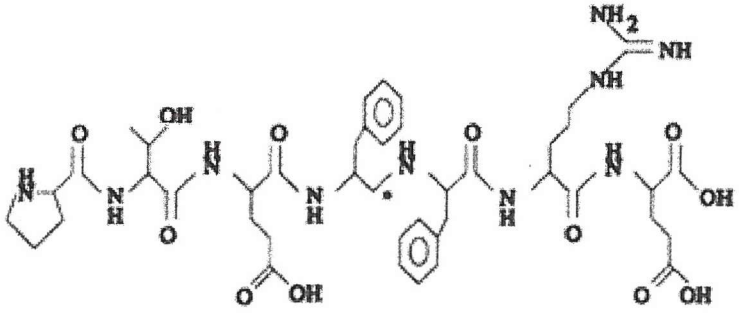


Figure 2.13 Showing the chemical composition of the H256 inhibitor, * indicates the inhibitory group.

Figure 2.14 Showing a protein crystal of endothiapepsin co-crystallised with H256.

PD-135,040 (Figure 2.15) is an inhibitor which contains two fluorine atoms whose electron withdrawing properties cause the inhibitor fluroketone group to readily hydrate into a geminal-diol form. The inhibitor has the sequence TSM-DPH-His-CHF-MOR and its K_i for endothiapepsin has not been measured. Where TSM is a tert-butylsulfonyl group, DPH is a deamino-methyl-phenylalanine, CHF is a cyclohexylfluorostatone and EMR an N-aminoethylmorpholine. The structure of this inhibitor bound to endothiapepsin at medium resolution (2.30 Å) has been detailed in Bailey and Cooper 1994.

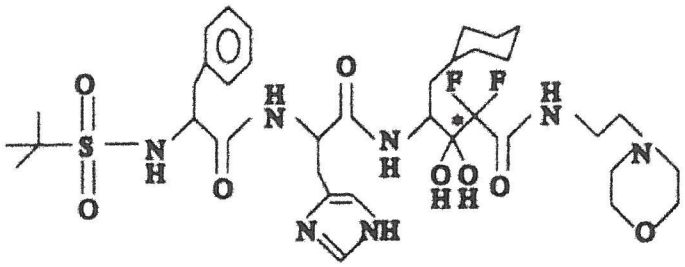


Figure 2.15 Showing the chemical composition of the PD-135,040 inhibitor, * indicates the inhibitory group.

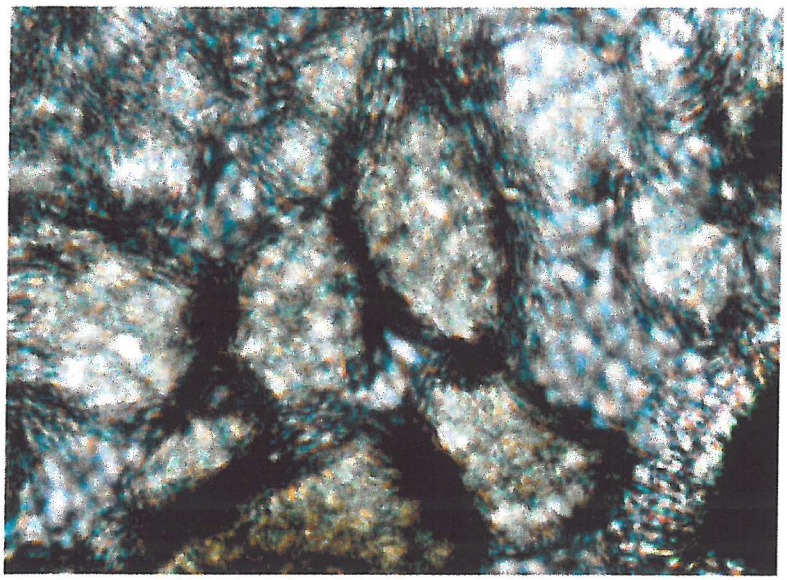


Figure 2.16 Showing a protein crystal of endothiapepsin co-crystallised with PD-135,040. These crystals were grown via the hanging drop method and a film had grown over the drop which could not easily be removed.

Chapter 3

Monochromatic X-ray Diffraction

Theory and Practice

Protein crystals

A crystal is a periodic arrangement of a motif in a lattice. The motif can be a single atom, a small molecule, a protein or any combination thereof. When the motif is repeated in three dimensions a simple crystal is formed. Very often the motif is also referred to as the 'asymmetric unit', which can be subjected to a number of symmetry operations yielding differently oriented copies. The unit cell of a crystal is the smallest and simplest volume element that is completely representative of the whole crystal. Roughly speaking protein crystallography requires a crystal of at least 50 μm in its shortest dimension. Protein crystals are held together by weak interactions such as hydrogen bonds between hydrated protein surfaces. Protein crystals are not a perfect array of unit cells rather they can be thought of as mosaics of many arrays in rough alignment with each other (Figure 3.00).

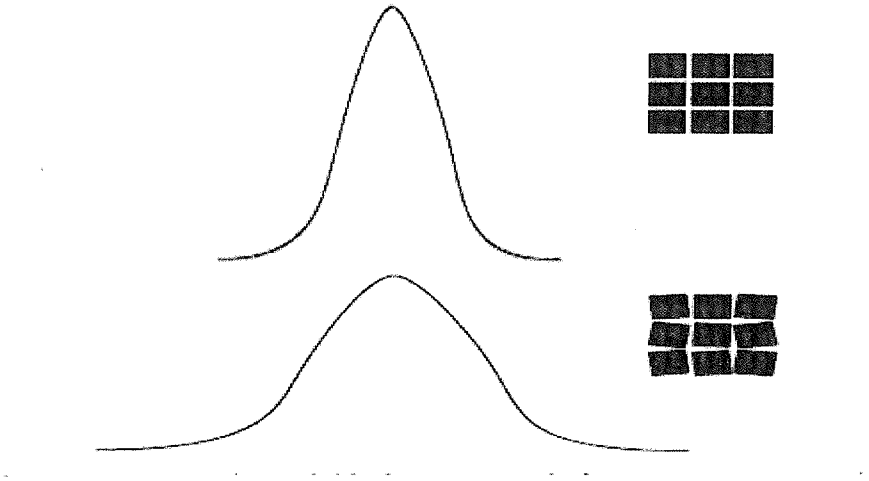


Figure 3.00 showing the effect of mosaicity on reflections, in the top diagram the regular spacing of unit cells gives a sharp peak. While in the bottom diagram the irregular spacing leads to a wider peak. (Figure from McRee 1999a)

The result of this mosaicity is that an X-ray reflection actually emerges from the crystal as a narrow cone rather than a perfectly linear beam. Thus the reflection must be measured over a very small angle rather than a single well defined angle. As the unit cells in the protein crystal are not all in a uniform orientation due to the weak interactions between protein molecules, the mosaicity increases. Thus the reflections from protein crystals suffer a greater mosaic spread than do those from more ordered crystals (McRee 1999a). Another factor besides mosaicity that causes blurring of reciprocal lattice points is the wavelength range present in the incoming beam as no beam is perfectly monochromatic. In practice there is always a range of wavelengths propagated through the crystal, which can be given by $d\lambda$. Furthermore, the incoming beam is never perfectly collimated, so the incident waves impinge on the crystal over a range of incident angles depending on the degree of divergence in the incoming beam.

Cryocrystallography

Protein crystals exposed to X-ray radiation sources can quickly suffer radiation damage. Therefore cryocooling is a method that can be used to alleviate the damage caused to the crystal by X-ray radiation. This serves three main purposes. It increases the lifetime of the crystal in the X-ray beam, allows data collection at higher resolutions and increases the signal-to-noise ratio. The process involves soaking a crystal in mother liquor containing a cyroprotectant followed by flash cooling to cryogenic temperatures (100 K). At this temperature the liquid surrounding the crystal is transformed into glassy solid, a process known as vitrification (Garman 1999). Cyroprotectants in common use include glycerol, ethylene glycol and PEG 400. The cyrocooling of crystals allows optimal conditions for their storage and transport to synchrotrons. However cryo-crystallography can sometimes lead to an increase in the mosaicity of the crystal and determination of the appropriate conditions can be difficult (Garman 1999).

The crystal lattice

Crystal lattices can be described mathematically by delta functions and convolutions. A delta function is non zero at one point only, i.e. 0 everywhere bar a single point where it has a finite value say when x=4. If we have a lot of different delta functions which peak at x=1, x=2, x=3 etc when they were added together then they will generate a lattice plane. A convolution is a mathematical way of multiplying and summing two functions. Given two functions f(x) and g(x), the convolution product is represented by fg⊗(u) where u is a variable which can take on the same values as x. The convolution is calculated by taking one of the functions, sliding it past the other, multiplying and summing them at each stage (Figure 3.01). Hence the crystal can be thought of as a convolution of the electron density (ρ) within the unit cell convoluted with the lattice, which is defined by the delta functions.

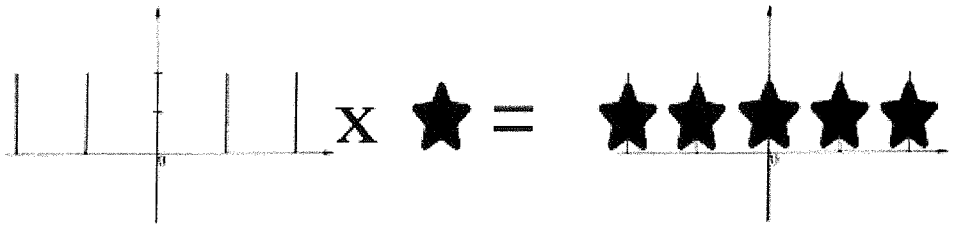


Figure 3.01 An example of a convolution. The star represents the unit cell while the peaks represent the delta functions which make up a lattice. The unit cell can be convoluted with the delta functions to form a crystal.

Many of the equations used in crystallography are Fourier transforms. These equations provide a way to calculate electron density from the diffraction pattern and vice versa. The properties of Fourier transforms will be covered later. Mathematics shows that the Fourier transform of a convolution is the product of two Fourier transforms. The Fourier transform of the crystal thus equals the Fourier transform of the electron density (ρ) multiplied by the Fourier transform of the crystal lattice as described by a three-dimensional delta functions. Hence to calculate the diffraction pattern of the crystal we need to calculate the Fourier transform of the crystal lattice. Mathematically the Fourier transform of a delta function is simply another delta function but the spacing of the peaks is different

(Figure 3.02). The peak separation in the Fourier transform of a delta function is proportional to the reciprocal of the peak separation in the crystal lattice. Hence the Fourier transform of the real crystal lattice is called the reciprocal lattice and the diffraction pattern is said to reside in reciprocal space.

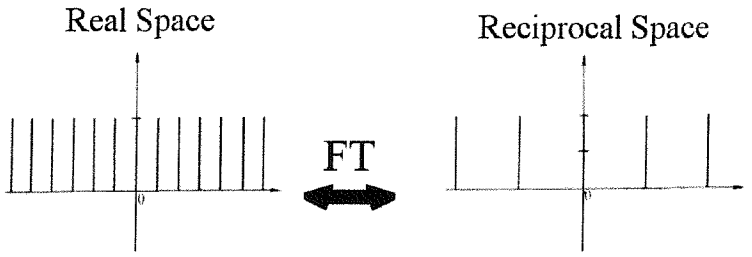


Figure 3.02 Showing the relationship between the real and the reciprocal lattice. The spacing between the peaks (delta functions) in the reciprocal lattice is 1/d, were d is spacing between peaks in real space. Diagram taken from Sherwood 1976.

The unit cell dimensions of the reciprocal lattice are given by reciprocal lattice vectors a*, b* and c* whose amplitudes are the reciprocals of a, b, c (the dimensions in real space) for an orthogonal unit cell. If the unit cell angles are not 90° they will be altered in the reciprocal unit cell. In real space the unit cell is defined by six values, the length of each edge is defined by a, b and c while the angle between the b and c edge is defined by the angle $\propto$. The angle β defines the angle between the a and c edges and γ defines the angle between the a and b edges (Figure 3.03).

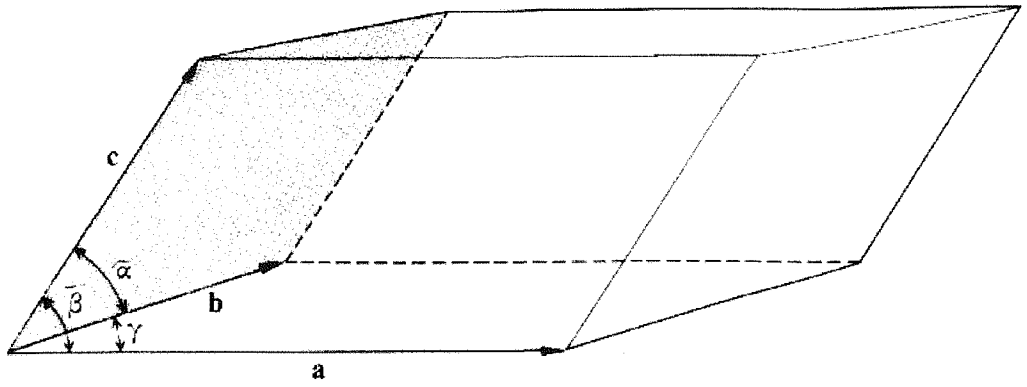


Figure 3.03 Showing a general three-dimensional unit cell with cell axes a, b and c and angles α, β and γ. Like all coordinate systems used in crystallography the system is right handed.

Depending on their a, b, c lengths and α, β, γ angle values, unit cells fall into seven crystal systems.

Crystal system	Unit cell constraints	Symmetry
Triclinic	Where a, b, c and α, β, γ can have any value	None
Monoclinic	a≠b≠c α=γ=90 β≠90	2-fold along the b axis
Orthorhombic	a≠b≠c α=β=γ=90	2 fold axes parallel to a, b, c
Tetragonal	a=b≠c α=β=γ=90	4 fold along c, 2 folds along a & b
Cubic	a=b=c α=β=γ=90	4 folds along a, b, c
Trigonal	a=b≠c α=β=90 γ=120	3 fold along c
Hexagonal	a=b≠c α=β=90 γ=120	6 fold along c

Table 1 Listing the seven different crystal systems with constraints and symmetry operators. The symbol ≠ means not necessarily equal.

A unit cell that only has lattice points at its corners is called a primitive lattice. Besides having lattice points in each corner in unit cells it is common for crystals to have an extra lattice point (motif) in their unit cells, these unit cells are called non primitive. There are three types of extra motifs, there is the internal centred cell (I) in which there is a extra point in the middle of the cell. The face centred cell (F) in which there are six extra lattice points one in the each face of the unit cell. And the (C) centred cell that has two extra lattice points one in opposite faces of the unit cell (Figure 3.04).

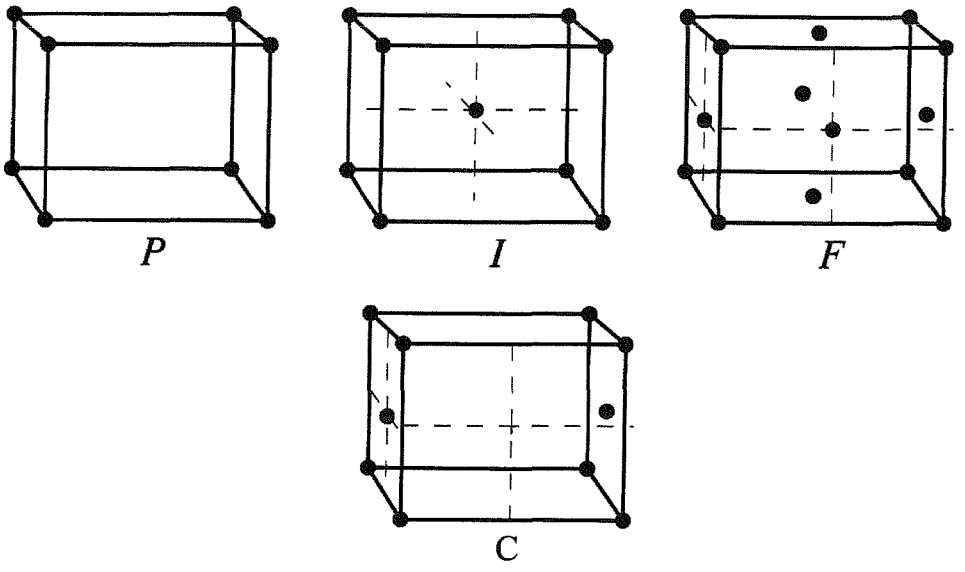


Figure 3.04 Showing the four possible types of unit cell (P) Primitive, (I) body centred, (F) face centred and (C) centred (adapted from Rhodes 2000).

So there are seven crystal systems and four types of centring leading to 28 possible lattices. However some of these possibilities are redundant, the net result being that there are 14 types of lattice called the Bravais lattices, seven of which are primitive and seven are not. All these lattices generally have symmetry axes coincident or parallel with the unit cell edges.

Space Groups

The crystallographic symmetry found within a unit cell is given by its space group, which details the relationship between equivalent positions in the unit cell. All symmetry elements in protein crystals are translations, rotations or screw axes which are rotations and translations combined. Translation simply means movement by a specified distance, usually by a fraction of a unit cell edge. In space group symbols, rotation axes are represented by an integer n e.g. a 4 fold rotation generates the same structure after a (360/n) or 90 degree rotation. The screw axis results from a combination of rotation and translation n_m where n is an n fold screw axis with a translation of m/n of the unit translation. Thus the screw axis or screw rotation is the combination of a rotation and a translation parallel to the rotation axis. For example a 2_1 axis consists of a rotation of 180 degrees about the 2_1 axis which is on or parallel to one of the unit cell edges. This is then followed by a translation or movement along the axis parallel to the 2_1 axis equal to half the length of the unit cell edge. A 3_2 axis which consists of a 120 degree rotation about one of the unit cell edges followed by a movement of two thirds of the length of the unit cell edge along an axis parallel to the unit cell edge. If a space group has a 2_1 axis along z then positions x, y, z and positions -x, -y, z+ ½ are symmetrical, where the z + ½ represents the ½ cell edge translation along z. Point groups are symmetry operations which act at a point in space leaving the point unaltered. Rotations, mirrors and roto-inversions (rotation + inversion) all act by leaving one point (the origin) unchanged. These operations combine to give 32 different point groups. Since the point group of a crystal is often reflected in its physical shape, the point groups are also referred to as crystal classes. Point group symmetry never involves translations e.g. the point group for 2_1 is 2. It is possible to combine all forms of symmetry element with the 14 Bravais lattices to derive all the possible symmetric arrangements of points or molecules in 3D space. There are 230 such arrangements called space groups. As proteins contain chiral amino acids, mirror and inversion symmetry cannot be adopted by proteins on crystallisation. This means there are only 65 allowed space groups for proteins. For proteins, orthorhombic and monoclinic space groups tend to be very common.

X-ray Generation

All of the X-ray diffraction data sets collected for this thesis were obtained at European synchrotron light sources (Desy, Hamburg & ESRF, Grenoble). Data were collected on the PD-130,328, CP-80,794, CP-129,541, H256 and PD-135,040 complexes at the ESRF synchrotron in Grenoble. While data on the H189 complex were collected at the EMBL outstation at the DORIS ring in Hamburg. A synchrotron radiation light source consists of an electron or positron storage ring, which holds high-energy electrons in a circular orbit via the use of bending magnets. As electrons are accelerated around the storage ring they emit synchrotron radiation when they are turned via the bending magnets (Figure 3.05). This radiation is continuous at the wavelengths used by protein crystallographers and is also at least two to three orders of magnitude as bright as radiation from the best rotating anode generators.

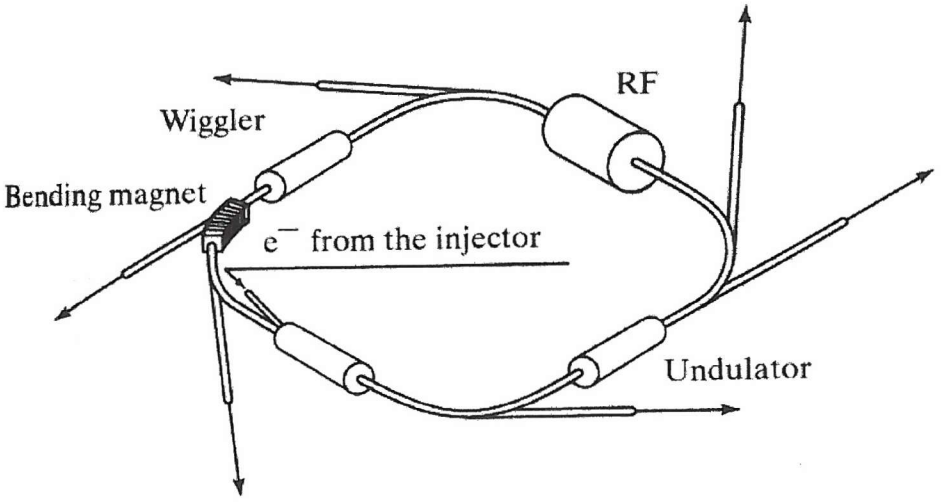


Figure 3.05 Showing the generalised layout of an electron storage ring used at a synchrotron. Bending magnets are responsible for turning of the electron beam while the RF units are used to move the electrons around the storage ring. The wiggler and undulator devices are used to increase the intensity of the X-ray beam. Diagram taken from Arndt 2001.

The brilliance of the synchrotron X-ray beam is often increased by the use of multipole magnet insertion devices such as wigglers and undulators which are placed in the straight sections of the storage ring and combine several beams from

local excursions of the electron path from the storage ring. The optics for the X-ray beam are also superior at synchrotron sources leading to a highly collimated high intensity beam, ideal for high resolution data collection. A monochromatic (single wavelength) source of X-rays is desirable for most crystallography because this fixes the diameter of the sphere of reflection to a single value meaning all reflections will be singlets.

CCD detectors

All X-ray data were collected using charged coupled device detectors more commonly known as CCDs which are at present the most useable and accurate large area detectors available for the measurement of the X-ray energies of interest to protein crystallographers. CCD detectors are formed from three main components: an energy converter (usually phosphor) to convert X-rays into photons of visible light, an optical relay (fibre optic or lens based) to carry the signals and de-magnify them, and a CCD chip (Figure 3.06).

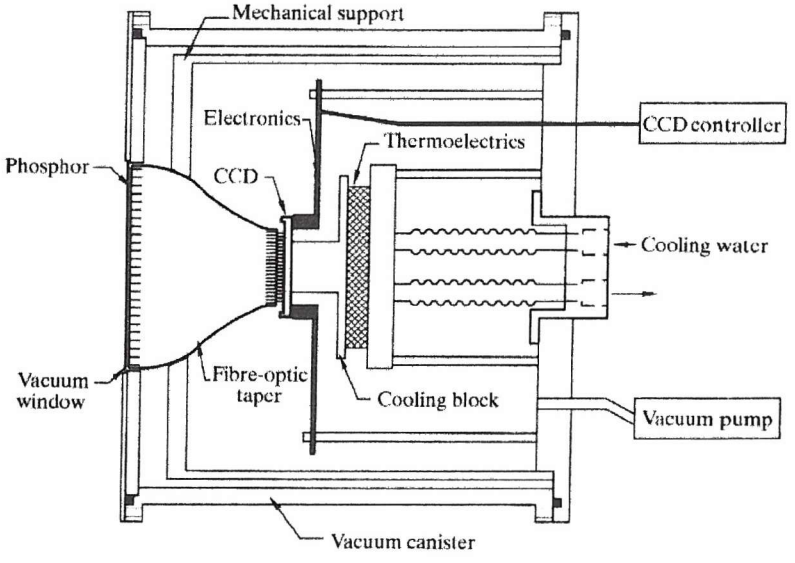


Figure 3.06 Showing the typical layout of a CCD based X-ray detector. The X-rays are converted into photons of visible light by the phosphor. These photons are then demagnified by the fibre optic taper onto the CCD where they are converted into an electric signal. Diagram taken from Gruner *et al* 2001.

Demagnification is required as scientific grade CCDs are only manufactured in relatively small sizes (25mm x 25mm) but as optical image reduction is inherently inefficient this demagnification is limited to a ratio of 4:1 without image intensification before demagnification. Light transmission in image reduction is given by

$$n.a. \times M^2$$

where n.a. is the numerical aperture of the optical system and M is the linear magnification factor. When used in image reduction lens based systems have n.a. values which typically give 2 % light transmission for a 3:1 reduction. In contrast fibre optic tapers have light transmission values of 13 % for the same reduction making them ideal to transmit the light given off by X-rays striking the phosphor. The only disadvantage of fibre optic tapers is their cost; they are around three times more expensive to produce than a lens.

In a CCD the visible light photons generated by the phosphor and demagnified by the lens or fibre optic taper are converted into charge carriers within the silicon of the CCD. A typical CCD connected to an optimised phosphor screen with a 3:1 fibre optic demagnification taper will give 10-30 recorded electrons per 10 keV (Gruner *et al* 2001). CCDs are normally cooled to well below room temperature to reduce thermally generated dark current that is a source of noise. The dark current drops by a factor of 2 for a temperature reduction of 5-7 K therefore the temperature of the CCD must be well regulated. Another source of dark current noise are surface defects on the CCD itself. To help minimise these effects, multiphase pinned (MPP) CCDs use charge implants within the pixel structure to move the charge collection area away from the surface of the CCD. A side effect of this is a reduction in the dynamic range of the pixel. Even with this limitation MPP CCDs are rapidly becoming the norm in most X-ray detectors. Multiple CCDs can be used in a single detector, which is the case in the ADSC Quantum R4 detector which utilises 4 CCDs to provide a large area detector suitable for protein crystallography. The MAR Research detector utilises a single CCD and

hence is a smaller detector than the Quantum 4 however it is significantly cheaper. Phosphorus is used as an energy converter because of its high atomic number necessary to make a thin screen increasing spatial resolution while maintaining a high X-ray stopping power.

X-ray diffraction

X-ray scattering is an interaction between electro-magnetic waves (X-rays) and the electrons in atoms. If an electromagnetic wave is incident on a system of electrons the electrical and magnetic components of the wave exert a force on the electrons. This causes the electrons to oscillate with the same frequency as the incident wave. The oscillating electrons act as radiation scatterer and they emit radiation of the same frequency as the incident radiation in all directions. The X-ray scattering amplitude depends on the number of electrons in the particular atom. The X-ray scattering amplitude decreases with increasing scattering angle (θ) and is higher for heavier atoms. One must realise though, that the scattering factor contains additional (complex) contributions from anomalous dispersion effects (essentially resonance absorption), which becomes substantial when the wavelength of the incident radiation approaches the X-ray absorption edge of the scattering atom. These anomalous contributions can be exploited in the MAD phasing technique. In X-ray crystallography, the diffracted beams are separately observed and their amplitudes can be measured from the intensities of reflections on a detector. These diffraction amplitudes are the Fourier transforms of the electron density. Although each atom in the unit cell of the crystal scatters X-rays in all directions only those scattered X-rays which travel in certain directions will interfere constructively to give diffraction spots. Every atom in the unit cell scatters X-rays with a phase that is dependent on its location in the unit cell and the directions of the incident and scattered X-ray beams. The amplitude of the diffracted beam is dependent on the number of electrons in the atom. Thus every reflection has its amplitude and phase affected by every atom in the unit cell.

Bragg's Law

Bragg showed that a set of parallel planes with indices h, k, l and interplanar spacing d_{hkl} produce a diffracted beam when X-rays of a given wavelength impinge on each plane at an angle θ. Constructive interference occurs only if θ meets the condition below.

$2d_{hkl}\sin\theta=n\lambda$ where n is an integer

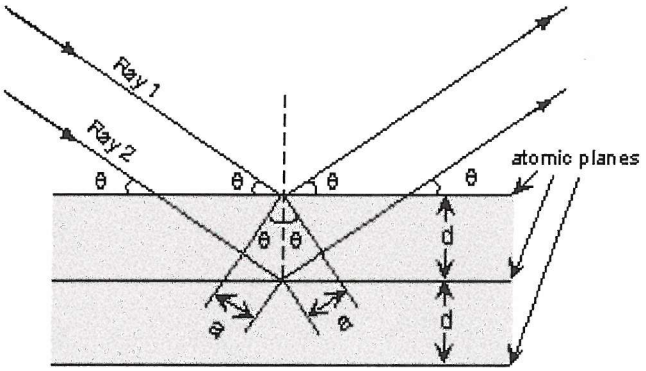


Figure 3.07 Showing diffraction from Bragg planes separated by spacing d. Constructive interference occurs the difference in path length (2a) is a multiple of the radiation wavelength.

If the difference in path length (shown as 2a in Figure 3.07) for rays reflected from successive planes is equal to an integral number of wavelengths of the impinging X-rays then rays reflected from successive planes re-enforce each other and produce a strong refection. Theta (θ) is the angle of diffraction and its sine is inversely related to interplanar spacing d_{hkl}. Each set of parallel planes in the crystal produces one reflection. The intensity of that reflection depends on the electron distribution projected onto the planes that produce the reflection.

Bragg's law in reciprocal space

In the reciprocal lattice the three reciprocal cell edges have length a*=1/a, b*=1/b, c*=1/c (for an orthogonal unit cell). If a sphere is drawn with a radius of 1/λ with the crystal at its centre (C) on the incoming X-ray beam an Ewald sphere is constructed (Figure 3.08).

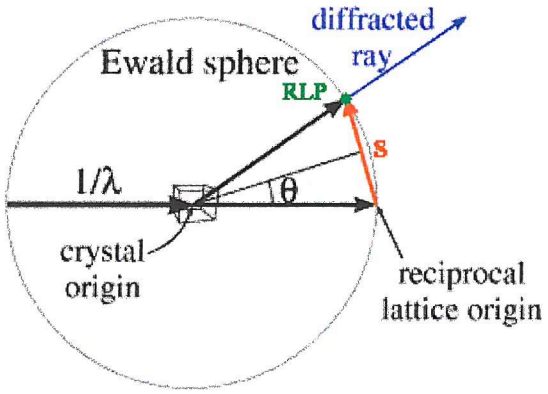


Figure 3.08 Showing the Ewald sphere, a construction for visualising diffraction. The diameter of the sphere is the reciprocal of the wavelength of the X-rays used. The construction has two origins, the crystal origin is the origin in real space while the point where the incident X-ray beam exits the sphere defines the origin of reciprocal space. Diffraction occurs when a RLP touches the Ewald sphere with the scattering vector **S** defining the RLP in reciprocal space.

When the crystal is rotated around its origin the reciprocal lattice rotates and different reciprocal lattice points (RLPs) come into contact with the edge of the Ewald sphere. Bragg's law is then satisfied and a reflection occurs from the direction of the circle centre (crystal origin) towards the reciprocal lattice point (RLP) whenever a reciprocal lattice point comes into contact with the sphere. The origin of the reciprocal lattice is defined as where the direct beam exits the reciprocal sphere. The scattering vector **S** gives the vector between the reciprocal lattice origin and the RLP that the reflected beam passes through. The amplitude of the scattering vector S is the inverse of the distance between the Bragg planes thus $S=1/d_{hkl}$. Constructive interference only occurs when $\underline{S}=h\underline{a}^* + k\underline{b}^* + l\underline{c}^*$, in this case all unit cells scatter in phase and h, k, l are always integers. Three indices h, k, l

identify a particular set of equivalent parallel planes. The index h is the number of planes per unit cell in the a direction, and likewise k and l refer to the number crossing b and c. All planes perpendicular to the ab plane have indices hk0 and likewise all planes perpendicular to the bc and ac planes have indices 0kl and h0l respectively. Indices can be negative as well as positive e.g. the (3,1,0) planes are the same as (-3,-1,0) planes but the (-3,-1,0) planes cross a and b in their negative directions. If a plane has 0 for any of its h, k, l values this means that the planes are perpendicular to that axis. Both the (h, k, l) and (-h,-k,-l) planes produce reflections of equal intensities but with opposite phases (Friedel's law). This reflection pair is called a Friedel pair or mate. The h, k, l indices also specify a vector in reciprocal space perpendicular to the reflection planes with a length equal to the spacing between reciprocal lattice planes. This is **S** the scattering vector. The amplitude of a particular structure factor F_{hkl} indicates the extent to which electron density is concentrated to planes parallel to the Bragg planes. While its phase indicates the position of planes of high electron density relative to the Bragg planes.

Real and Reciprocal space

In real space the variable **r** represents a three-dimensional spatial coordinate. This point is defined by the Cartesian coordinates x, y, z. Any value of **r** will correspond to a particular point in space, and the complete range of the values of **r** will define all space. The distances x, y, z used to define **r** in real space are given in normal units of length. Mathematically any variable may be thought of as defining a space such as the scattering vector **S** that is defined by three components h, k, l each of which has values of length $^{-1}$. Thus we can set up an analogue of the three-dimensional Cartesian coordinate with each h, k, l value representing a component of the scattering vector **S**. As **S** takes on different values, we may identify points relative to the coordinate system using any value of **S**. The space defined by the variable **S** is called k space, reciprocal space or Fourier space and is defined in units of reciprocal length. The Fourier transform

enables coordinates in real space to be transformed into reciprocal space and vice versa.

Diffraction corresponding to a particular RLP occurs whenever the point intersects the Ewald sphere. But as the RLP is smeared, the intersection of any given point with the Ewald sphere is not a precise event that occurs in a single incident. To allow for this the protein crystal is rotated about an axis with a constant angular velocity ω, consequently the reciprocal lattice rotates with the same constant angular velocity ω about its origin. Any RLP therefore sweeps through the Ewald sphere with diffraction occurring over a range of values corresponding to the amount of smearing of the RLP. The relationship between intensity profiles for an ideal RLP and an actual RLP is shown in Figure 3.09. Since protein crystals give diffraction spots of a diffuse nature, the intensity of any RLP is proportional to the area under the intensity curve, which is called the integrated reflecting power.

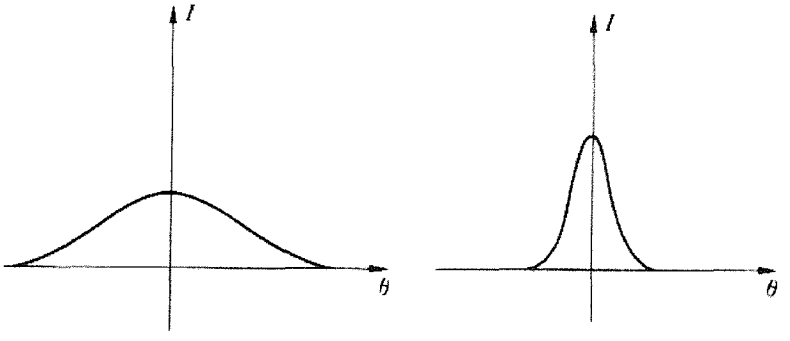


Figure 3.09 Showing the intensity profiles of two different Bragg reflections, as the intensity of the RLP is proportional to the area under the curve (a) is the stronger reflection. Diagram taken from Sherwood 1976.

The Lorentz factor

As a RLP passes through the Ewald sphere the area of intersection of the surface of the sphere with the RLP will vary according to the shape of the RLP and the precise geometry of the intersection. This is shown in Figure 3.10.

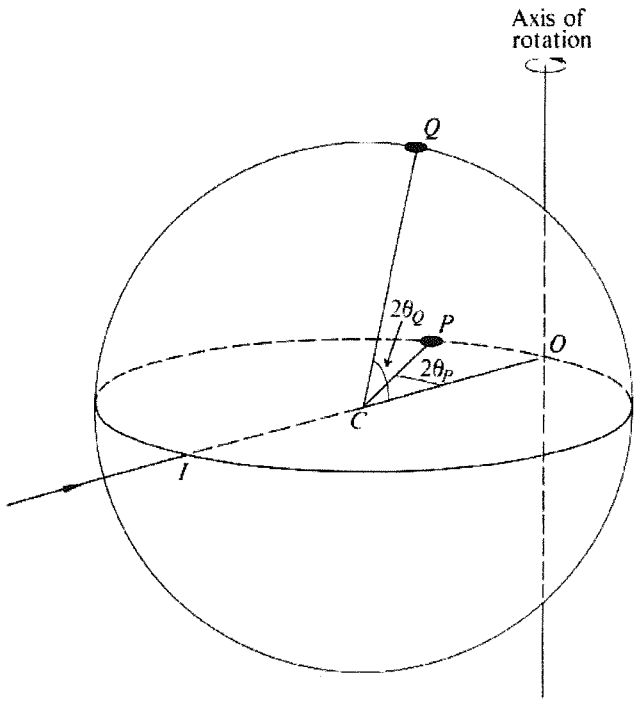


Figure 3.10 Both P and Q are RLPs on the surface of the Ewald sphere with O being the origin of reciprocal space. However due to the Lorentz factor they spend different amounts of time in a reflecting position. Diagram taken from Sherwood 1976.

In Figure 3.10 there are two RLPs (P and Q) which have the same shape. Point P intersects with a diametral plane of the Ewald sphere, whereas point Q intersects the Ewald sphere at an upper level. The passage of point P through the Ewald sphere as the crystal rotates is a fairly sharp event. But as point Q passes through the Ewald sphere at a glancing angle its passage through the Ewald sphere is extended compared to point P, thus the opportunity for the diffraction events from points P and Q are quite different. In addition to the above geometrical effect there is a further factor effecting diffraction from points P and Q. The closer a RLP is to the rotation axis the lower the linear velocity will be, thus point Q will pass through the Ewald sphere more quickly than point P.

Absorption

As an X-ray beam passes through any material, the electric field causes electronic excitations that increase the thermal energy of the electrons and theoretically the temperature of the material will rise. Therefore the thermal energy of a protein crystal will rise at the expense of energy from the incident X-ray beam, thus the total energy of the beam is greater on entering the crystal than leaving it. This effect is known as absorption and represents possibly the largest source of un-correctable errors in an X-ray data set. Experiments have shown that the reduction in intensity of the incident beam (-dI) is proportional to the product of the distance through which the beam travels (dx) and the local intensity (I).

$$-dI \propto Idx$$

The constant of proportionality is a property of the material of the specimen and is called the linear absorption coefficient (μ) (Sherwood 1976). Therefore

$$-dI = \mu Idx$$

or

$$I(x) = I_o e^{-\mu x}$$

Which I(x) is the intensity of the X-ray beam at any distance x within the crystal, and I_o is the incident intensity. The linear absorption coefficient varies with wavelength in an erratic manner (Figure 3.11).

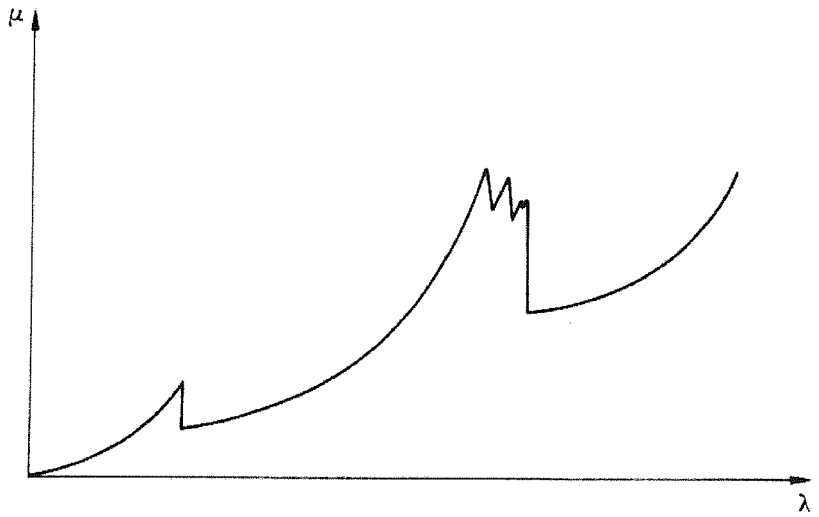


Figure 3.11 Showing the variation of the linear absorption coefficient (μ) with wavelength (λ). The peaks in absorption to changes in the quantum state of electrons within an element. Diagram taken from Sherwood (1976).

The sharp discontinuities in the behaviour of the linear absorption coefficient (μ) are called absorption edges. The wavelengths at which they occur are related to those X-ray energies which cause changes in the quantum states of electrons within a single type of atom. These absorption edges are often utilised in anomalous scattering techniques and are also significant in the design of X-ray filters. The correction for absorption is a very difficult matter as X-rays passing through different parts of the crystal have different path lengths within the crystal (Figure 3.12).

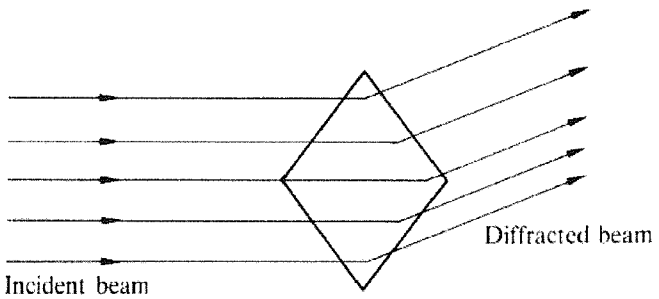


Figure 3.12 The path length through the crystal for the incident beam is not uniform. This leads to problems in correcting for absorption by the crystal. The diffracted beams also travel different distances through to reach the detector. (diagram taken from Sherwood 1976)

This causes a large problem, as the shape and size of protein crystals are extremely variable. Therefore an explicit absorption correction would be difficult to determine. However the effects of absorption are corrected for in data scaling. The level of absorption also varies with the path length of the diffracted X-rays through air. Thus the path length for each reflection differs. Higher resolution reflections will therefore be associated with increasing absorption compared to the lower resolution reflections. Modern attempts to correct for absorption involve a least squares fit between the differences of symmetry related reflections as a function of some parameter believed to be a function of absorption.

The number of measurable reflections

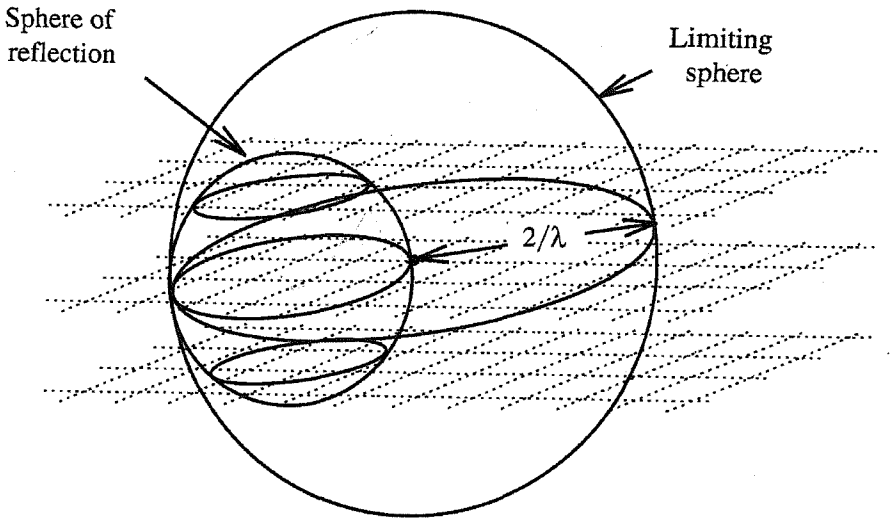


Figure 3.13 The sphere of reflection has a radius of $1/\lambda$ thus any RLP within $2/\lambda$ of the origin can be rotated into contact with the sphere of reflection. The limiting sphere thus has a radius of $2/\lambda$ about the origin of reciprocal space. Diagram taken from Rhodes 2000.

The number of RLPs within the limiting sphere is equal to the maximum number of reflections that can be produced by rotating the crystal through all possible orientations in the X-ray beam. This shows that the unit cell dimensions and wavelength of the X-rays determine the number of measurable reflections. Shorter

wavelengths make a larger sphere of reflection increasing the maximum number of reflections. Large unit cells give smaller reciprocal unit cells whereas small unit cells give large reciprocal unit cells. Thus large unit cells increase the maximum number of reflections that have to be measured.

Structure factors

A structure factor describes one diffracted X-ray beam, which produces one reflection received at the detector. A structure factor represents the resultant X-ray scattering power of the whole crystal structure. Since the whole structure consists of a large number of unit cells all scattering in phase with each other, the resultant scattering power is actually calculated for the contents of one unit cell only. The structure factor therefore represents the resultant amplitude and phase of scattering of all the electron density distribution of one unit cell. The amplitude is calculated relative to the amplitude of scattering from an isolated electron. The phase is calculated relative to a phase of zero for hypothetical scattering by a point at the origin of the unit cell. The resultant is calculated as a superimposition of waves, one from each atom in the unit cell, each wave having an amplitude which depends on the number of electrons in the atom and a phase which depends on the position of the atom in the unit cell. Structure factors can be represented as complex vectors where the length of the vector represents the amplitude of the structure factor F_{hkl}, the amplitude being related to the number of electrons in the atoms that form the diffracted wave and their positions. The amplitude of scattering also depends on the mobility of the atoms in the crystal. Atoms that are thermally mobile will scatter X-rays more weakly than stationary atoms. The phase of the structure factor is represented by the angle α that the vector makes with the real axis (x axis) of the Argand diagram and is given in radians. The phase of a structure factor tells us the position of the vector at some arbitrary origin. In Friedels law $I_{hkl}=I_{-h-k-l}$ however F_{hkl} and F_{-h-k-l} are equal in amplitude but not equal in phase. The structure factors of Friedel pairs have the same amplitude but have opposite phases thus F_{-h-k-l} is the mirror image of F_{hkl} with the real axis serving as a

mirror. Because the diffractive contributions of atoms are additive each contribution can be represented as a complex vector.

If F represents a structure factor of a three atom structure in which f1, f2 and f3 are atomic scatting factors, f represents the amplitude and its angle phase α of each wavelet. The vector sum F=f1+f2+f3 is obtained by placing the tail of f1 at the origin the tail of f2 on the head of f1 and so on while maintaining the phase angle of each vector. The structure factor F is then given, as it is a vector with its tail at the origin and its head touching the head of the final vector (Figure 3.14). This process sums up amplitudes and phases so the resultant length of F represents its amplitude and the resultant angle alpha its phase angle (the atomic vectors may be added in any order with the same result).

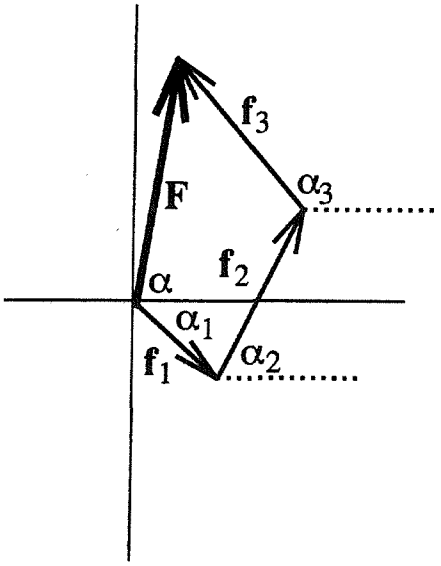


Figure 3.14 F is formed from the summation of amplitudes and phases from each f, where F is the structure factor and each f represents the contribution from a single atom. The order in which summation of the f values takes place does not affect the value of F. Diagram taken from Rhodes 2000.

The structure factor F can also be described by a complex number

$F = a + ib$

where a is defined as the value of the real axis and ib is the value on the imaginary axis. However the complex conjugate of F written as F* is defined as

$F^* = a - ib$

The distinction between a complex number and complex conjugate is solely due to a change in sign of the imaginary component. For a complex number F and F* are the mirror image of each other along the real axis. The multiplication of a complex number with its complex conjugate gives

$FF^* = |F|^2$

Which is a real number equal to $a^2 + b^2$ and can be measured in a physical experiment (Sherwood 1976). However since

$|F|^2 = a^2 + b^2$

there is always an ambiguity in the sign of the imaginary value. This problem exists for all complex numbers in crystallography and it leads to a phase ambiguity in which there are two correct values for F.

Fourier Transforms

Both the electron density and structure factor equations are Fourier transforms. Fourier transforms have a number of useful properties, the primary one being that a Fourier transform is its own inverse. Thus if you apply a Fourier transform twice you get the original function back, meaning that as the structure factors are the Fourier transform of the electron density the Fourier transform of the structure

factors is the electron density. Fourier showed that for any function f(x) there exists another function F(h) which is the Fourier transform of f(x) and the units of the variable h given in the formula below are the reciprocals of the units x.

$$F(h) = \int f(x) e^{2\pi i(hx)} dx$$

As the electron density represented by ρ(xyz) is a three dimensional wave its Fourier transform has F_{hkl} terms. In the diffraction pattern high frequency Fourier terms occur at the outside edge of the diffraction pattern and give fine details and have high indices. Low frequency Fourier terms that give gross details and have low indices appear towards the middle of the diffraction pattern. A Fourier series is a set of functions described by the sum of simple sine and cosine functions where their wavelengths are integral fractions of the wavelength of the complicated function. The Fourier series that describes a diffracted X-ray is called a structure factor equation

$$F_{hkl} = \int_x \int_y \int_z \rho(x,y,z) e^{2\pi i(hx+ky+lz)} dxdydz$$

Each diffracted X-ray beam that arrives at the detector to produce a recorded reflection can be described by a Fourier series. The computed sum of the series for the reflection h, k, l is called the structure factor F_{hkl}. Another form of the structure factor equation has a Fourier term for every atom in the unit cell. This equation states that each reflection is the result of diffractive contributions from all atoms in the unit cell. Thus the structure factor is a wave created by the superposition of many individual waves each resulting from diffraction by an individual atom.

The Electron density Equation

The view of the ρ(x, y, z) as the Fourier transform of the structure factors requires the amplitude, frequency and phase of each reflection.

$$\rho(xyz) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l} F_{hkl}\, e^{-2\Pi i(hx+ky+lz)}$$

The intensity I_{hkl} of a reflection gives the reflection's amplitude by the equation $I(h, k, l)=|F(h, k, l)|^2$ or the amplitude of the structure factor can be calculated from:

$$|F| = \sqrt{I_{(hkl)}}$$

The h, k, l position specifies the frequency for that term, but the phase of each reflection is not recorded on the image plate. To calculate the electron density this value is required and its computation is referred to as the phase problem. In the electron density equation one has to divide by the volume (V) of the unit cell in order to calculate the electron density on the correct scale. As the electron density equation is written here it requires a sum over all structure factors in reciprocal space. Because of the resolution limit of the crystal this will generally be a sphere in reciprocal space. The space group of the protein and Friedel's law also act to lower the number of structure factors that need to be measured.

X-ray Data collection

In a monochromatic X-ray diffraction experiment only reflections whose RLPs happen to lie on or cross the Ewald sphere are recorded. In order to measure the intensities of additional reflections the crystal must be rotated with respect to the X-ray beam. This normally involves turning the crystal around an axis perpendicular to the X-ray beam. From the diffraction pattern the h, k, l position and the intensity of a reflection can be worked out. The largest reflection is at the origin corresponding to $F_{(000)}$, which is the sum of all electrons in the unit cell.

This reflection is at zero diffraction angle, i.e. in the primary beam path and not observable. The aim of a data collection strategy is to collect a complete dataset, i.e. one covering more than 95 % of the theoretically possible data points. Multiple observations of the same reflection are merged and give rise to the internal R-value which has many guises, R_{int}, R_{sym} or $R_{merge.}$ Often in monochromatic diffraction the crystal is rotated a set number of degrees about ϕ during each exposure. The main purpose of this is to ensure that each spot passes through the Ewald sphere. Care must be taken so that spatial overlap of reflections does not take place, which could occur if the oscillation angle is too great. The amount of background is also related to the oscillation angle, the greater the angle the greater the level of background on the diffraction image due to the inherent requirement for a longer exposure time. Oscillation of the crystal during data collection also affects the types of reflections observed in the diffraction pattern. As protein crystals are mosaics of submicroscopic crystals a RLP can be thought of as an ovoid rather than an infinitesimal point. This means that RLPs diffract over a very small ϕ range; diffraction is weak at the start of the ovoid then peaks at the middle and then lowers off towards the end. This fact means that for some of the reflections in the diffraction pattern the whole RLP is not measured on a single image. In other words the oscillation angle may not fully sample reflections diffracting at the edge of the start and end oscillation values. These reflections are called partials and they are found at the edges of diffraction lunes. They can be summed across diffraction images after scaling and included in the data set. They are also used for post refinement of the cell parameters and orientation matrix.

Symmetry and data collection

Only the contents of the asymmetric unit (or motif) are needed to construct the entire unit cell by applying the space group symmetry operations. Likewise only the contents of the asymmetric unit in reciprocal space are needed to reconstruct the entire diffraction pattern and hence the electron density of the asymmetric unit. The reciprocal space asymmetric unit is defined by the point group symmetry plus an inversion centre which is the so called Laue group. This defines the point

symmetry of the reciprocal lattice. In the absence of anomalous scattering, the reciprocal lattice is centrosymmetric and so is the diffraction pattern. The completeness of a data set is usually reported as the percentage of observed data (not using any intensity based cut off) compared to the total possible data in the asymmetric unit of the reciprocal space. Unique reflections often come from multiple observations of the same symmetry related reflections, which are merged into one unique intensity value. Extra symmetry in the diffraction pattern arises from space group symmetry elements. For example a two-fold axis along z will give rise to four reflections: (h, k, l) (-h, -k, l) (h, k, -l) and (-h, -k, -l) all with equal intensity. Although the screw character of a two-fold screw axis is not detected in the symmetry of the diffraction pattern it has an effect on the reflections along the corresponding 2 fold axis in the diffraction pattern.

Systematic absences

All operations involving translations such as screw axes yield observable extinctions (systematically absent reflections) in the diffraction pattern. Systematic absences in the diffraction pattern reveal symmetry elements in the unit cell. For example the $P2_1$ space group has the limiting condition 0k0: k=2n, meaning that reflections with h and l values of 0 and odd k values are systematically absent. Those with even k values are not extinguished; the reason for this is shown in Figure 3.15.

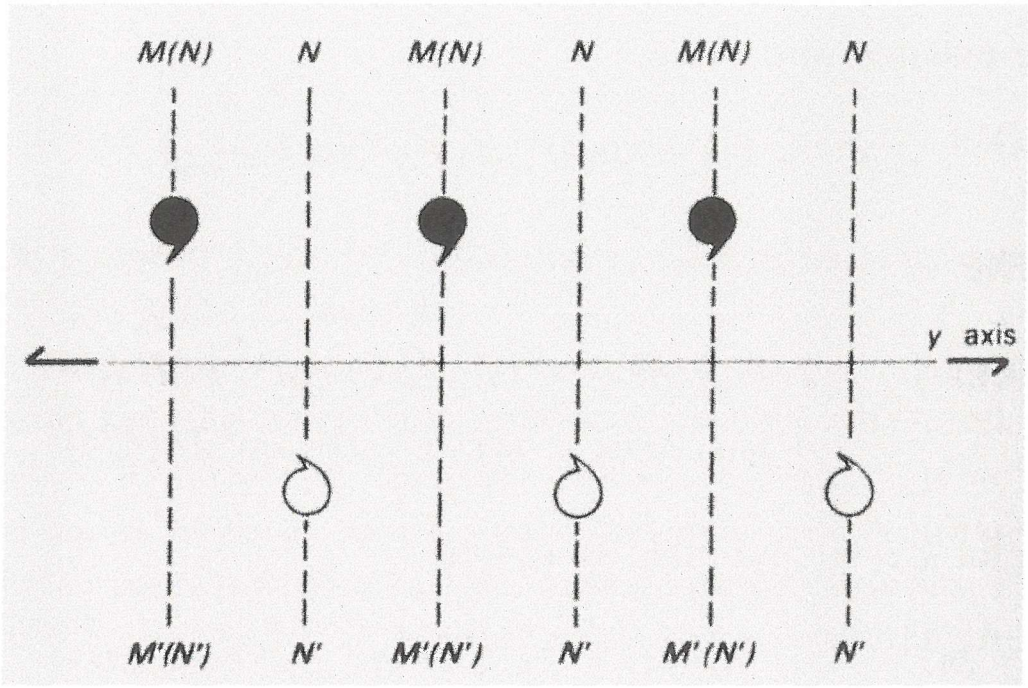


Figure 3.15 In a 2_1 axis d(N,N')=d(M,M')/2 thus the M,M' planes are halved by the N,N' family. (diagram taken from Ladd and Palmer 1993)

In the illustration of the 2_1 symmetry pattern the black motif represents a structure at height Z and the white motif represents the structure at height –Z after operating on it with the 2_1 axis. M,M' represents a family of planes (0,k,0) and N,N' a family of (0,2k,0) planes. Reflections from the M,M' planes are cancelled by the reflections from the N,N' planes because their phase change relative to M,M' is 180° for all odd axial reflections. For even axial reflections this phase difference is 360° i.e. constructive interference occurs. Put another way if the unit cell contains a twofold screw axis along the c edge then every atom in the unit cell is paired with a symmetry related atom that cancels its contributions to all odd numbered 00l reflections (Rhodes 2000).

Atomic displacement

Atomic displacement factors or B factors come into play during X-ray scattering, since they relate to the fact that atoms in the unit cell are not completely still. Atoms exhibit thermal motion and thus X-rays do not meet identical atoms in exactly the same position in every unit cell. This movement increases as the temperature increases and results in less well defined electron density in the final map. The amount of smear increases with temperature, which diminishes the scattered X-ray intensity especially at high scattering angles owing to the finite size of the electron cloud around the nucleus. For a given number of electrons the larger this cloud is and the more rapidly the atoms scattering power falls off with scattering angle (Merritt 1999). To account for this an angle dependent term can be added to the structure factor calculation thus

$$f = f_O \exp(-2\pi^2 \langle u^2 \rangle h^T h) = f_O \exp[-8\pi^2 \langle u^2 \rangle (\sin^2 \theta / \lambda^2)]$$

where $\langle u^2 \rangle$ is the mean square amplitude of vibration of the atom, h is reciprocal lattice vector, θ is the corresponding scattering angle and λ is the X-ray wavelength. The electron cloud of a vibrating atom averaged over time is larger than that of a similar stationary atom. As the magnitude of vibration is correlated with temperature, the parameter u is often called a temperature factor. However this is misleading since the smearing of the electron cloud at an atomic site in the crystal is a consequence of not only thermal motion, but also of stochastic variation in the true location of the atomic centre from one unit cell to the next (Merritt 1999). The vibration of an atom in a reflecting plane h, k, l has no effect on the intensity of the reflection h, k, l. Atoms in a plane diffract in phase and therefore a displacement in that plane has no effect on the scattered intensity. However the component of the vibration perpendicular to the reflecting plane does have an effect. In the simple case in which the components of vibration are the same in all directions the vibration is called isotropic. In this case the parameter $B=8\pi^2\langle u^2 \rangle$ (e.g. if $B=20\text{Å}^2$ then $u=0.25\text{Å}^2$) is used to describe an electron cloud which is uniformly smearing in all directions, thus the parameter u is a single

number giving an isotropic atomic displacement parameter. Anisotropic vibration is where the displacement is not the same in all directions. This motion is described by a series of ADPs (anisotropic displacement parameters). In this formulation the motion is described by an ellipsoid that can be oriented in any direction (McRee 1999a). An ellipsoid is described by a symmetric 3 x 3 symmetrical tensor matrix.

$$U = \begin{vmatrix} U_{11} & U_{12} & U_{13} \\ U_{21} & U_{22} & U_{23} \\ U_{31} & U_{32} & U_{33} \end{vmatrix}$$

but as the matrix is symmetric the lower three elements below the diagonal can be removed leaving six elements. The elements on the diagonal U_{11} U_{22} U_{33} specify the magnitude of the movement on the three axes while the off diagonal elements U_{12} U_{13} U_{23} specify the rotation of the ellipsoid off the principal axes. The anisotropy of atom (A) can be defined as the ratio U_{min} / U_{max} of the diagonal elements U_{11} U_{22} U_{33}. A spherical isotropic atom would have an anisotropy value of 1. In a survey of anisotropic thermal parameters the average anisotropy of proteins was found to be 0.45 for proteins refined at atomic resolution, with a number of proteins having an anisotropy around 0.54 (Merritt 1999a). The sample size used in this study was limited by the fairly low number of atomic resolution structures available making it hard to generalise the findings.

X-ray data processing

After the collection of the diffraction images the data is reduced to a series of h, k, l indices and observed intensities (I) and the associated error $\sigma(I)$. The main program used for this was MOSFLM (Leslie 1992) which is able to autoindex diffraction images. The first step in interactive autoindexing is to locate the positions of the diffraction spots using the 'find spots' option. This option finds

and displays the spots that will be used for autoindexing. The spots are selected if their $I/\sigma(I)$ values are greater than a certain minimum cut off which can be varied. The auto indexing routine then determines a number of values for different possible unit cell dimensions and space groups and lists a penalty value for each. The lower the penalty, the better the observed spots fit the particular unit cell and space group. After picking a set of values for the known unit cell and space group the predicted spot locations are displayed. A matrix file is also written to disc which describes the orientation of the unit cell in the beam as well as the unit cell dimensions. The spot positions are determined by the unit cell and space group. Only the amplitude of each spot gives any information on electron density. Thus it is critical that the predicated spot positions are as close a match as possible to the actual spot positions in order to measure their intensities as accurately as possible. One option to maximise this agreement is to run the refine cell option within MOSFLM. This refines the unit cell dimensions, crystal orientation as well as the position of the X-ray beam against a number of image segments. These segments consist of two or more diffraction images. Often two or more segments from different ϕ ranges of the data collection are used to refine the parameters. Full reflections are recorded within a single diffraction image. It is the partial reflections that are the most use in the postrefinement of the unit cell parameters particularly those that are recorded over two images (this is the reason that a segment must contain at least two images). However it is important to have a realistic estimate of the mosaic spread before refining the cell parameters. Versions 6.10+ of MOSFLM can estimate the mosaic spread during interactive processing. This is best carried out when there is a good match between the predicted and actual spot positions. The predicted pattern can be seen using the predict spots option which should hopefully predict the spot positions accurately. The refined matrix file along with a command file containing a list of MOSFLM parameters can then be used to integrate all the images in a data set. This is normally done in batch mode since it does not require any user interaction and produces a number of output files. The most important of these is the MTZ file that contains the actual intensities from all of the recorded h, k, l indices, and the summary file, which lists a number of statistics for each image that has been integrated.

Fourier synthesis density maps

Classically two types of Fourier synthesis $2F_o$-F_c,α_{calc} and F_o-F_c, α_{calc} are used to assess electron density for the current macromolecular model. The F_o-F_c map in which the calculated amplitude is subtracted from the observed amplitude is especially useful for finding corrections to the current model such as missing water molecules, alternative conformations and misfitted sidechains. The $2F_o$-F_c Fourier synthesis is the sum of a F_o map and a F_o-F_c map, it has electron density for the protein with additional features for the missing or incorrectly fitted features. Both of these classical Fourier synthesis are easy to interpret because it looks like protein density. However both these synthesis are heavily dependent on the quality of the calculated phases (α_{calc}) especially the $2F_o$-F_c synthesis (McRee 1999a); these syntheses are said to be phase biased. One way to try and minimise this phase bias is by the use of the σ_A weighted Fourier synthesis $2mF_o$-DF_c, α_{calc} and mF_o-DF_c, α_{calc} (Read 1986) in which two coefficients m and D are derived from a statistical analysis of the data. These coefficients vary between 0 and 1, with the aim of the analysis being to weight the terms by taking into account the differences between F_o and F_c.

Refinement

There are a number of computer programs designed to refine protein structures using different mathematical techniques. However the aim of all these programs is the same, to adjust the parameters of the model so as the structure factor amplitudes it generates (F_c) are a closer match to the observed structure factor amplitudes (F_o).

The most common way of doing this is by least squares methods in which the best set of model parameters is that which minimises the sum of the weighted squares of the differences between F_o and F_c.

$$\Sigma_{hkl} W_{hkl} (|F_o|-|F_c|)^2$$

Because the equations used are non-linear an exact solution is not possible and an approximate solution is generated in successive iterations. After each refinement cycle a new set of values for F_c are computed and shifts in the atomic parameters calculated until there are no significant shifts from the previous cycle. This indicates that the function has reached its minimum and the refinement has converged to the final parameter set. The parameters adjusted in a refinement cycle can include the (x, y, z) coordinates of the atoms in the structure, the occupancy as well as a B factor for each atom. The B factor can either be defined by a single number indicating thermal motion is the same in all directions (isotropic), or by an anisotropic B factor in which the thermal motion in each direction can differ. Anisotropic B factors give a more realistic model of thermal motion however they increase the number of observations per atom to nine compared to four for isotropic B factors. This becomes very important when calculating the observations to parameters ratio (observations/parameters). To calculate anisotropic B factors at least 2.25 times more observations are needed than to calculate isotropic B factors. For medium resolution structures (2.5 Å) this would increase the number of parameters above the number of observations making anisotropic B factor refinement impossible. An observations to parameters ratio of at least 2:1 is preferable for least squares calculations and the observations to parameters ratio should be as high as possible in refinement. Thus anisotropic B-factor refinement is only possible with high resolution data which inherently has more reflections giving more observations. One way to increase the data to parameter ratio is by the use of the known rules of stereochemistry for amino acids during refinement, which can be done in two ways. The first method is constrained refinement in which stereochemical information fixes bond lengths and bond angles within the model to set values. This reduces the number of parameters but

also reduces the number of degrees of freedom, so the structure may not reach the best position. The second method is restrained refinement which allows the bond angles and lengths to vary around a target value. Non crystallographic symmetry (NCS) which arises from symmetry within the asymmetric unit can also be used as a constraint or restraint to lower the number of parameters. In TLS refinement three matrices (Translation, Libration and Screw) are used to describe the atomic motion of a rigid body. This is useful for modelling anisotropic displacement parameters (ADPs) where atomic resolution diffraction data is not available. For each TLS group only 20 extra parameters are added to the refinement and a single ADP is calculated for the entire group.

Atomic resolution data refinement

Five of the six X-ray data sets collected on endothiapepsin with the different inhibitors were suitable for atomic resolution refinement (D_{min} < 1.2Å). The main aim of this was to conduct unrestrained refinement of the X-ray data. The higher the resolution of the data set the more reflections it contains, giving a more favourable observations to parameters ratio allowing ADPs to be refined. Unrestrained refinement is where stereochemical restraints are not used and the shifts are driven by the reflection data alone. This type of refinement, which is only possible at high resolution was a primary aim for work on the X-ray data sets of all endothiapepsin inhibitor complexes. Any differences in the carboxyl C-O bond lengths should be observable with atomic resolution data (Figure 3.16).

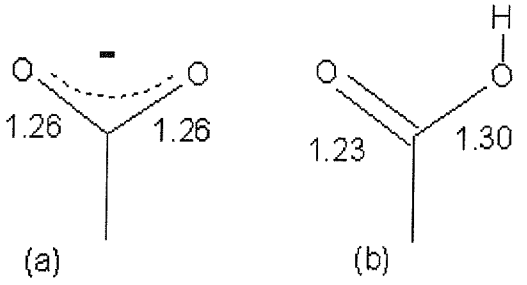


Figure 3.16 Bond length differences between unprotonated and protonated aspartates in Å. The carboxyl bond lengths in a negatively charged aspartate (a) would be expected to be equal while in a protonated aspartate (b) they would be asymmetric.

These bond lengths can therefore discern between protonated and negatively charged aspartates. The independent refinement of these carboxyl bond lengths could indicate which of the catalytic aspartates of endothiapepsin is protonated when the transition state inhibitor is bound. The low errors or ESDs (estimated standard deviations) on atomic coordinates in atomic resolution structures for well ordered parts of the molecule are typically around 0.01 Å at 1.0 Å resolution (Table 2) which enables meaningful bond length analysis (McRee 1999). The average atomic positional ESD for a 1.0 Å structure is around 0.03 Å. However the role of the B factors in estimating each bond length must be taken into account, as high B factors can reduce the certainty of the bond lengths.

Resolution	Average Atom Postional uncertainty
2.0 Å	0.32 Å
1.6 Å	0.13 Å
1.0 Å	0.03 Å

Table 2 A table showing the average atom positional uncertainties at a range of different resolutions.

At high resolution it is also possible to refine the occupancy of atoms in side chains where more than one conformation is visible in the Fourier maps. Split side chains are normally detected using $2mF_o$-DF_c and mF_o-DF_c density maps where extra density around a side chain indicates two or more possible conformations for the side chain. With atomic resolution data riding hydrogens can also be added to the model which means that the geometry of these hydrogens is fixed relative to the heavier atom to which they are attached. Thus they are not free to refine but move with the heavier atom to which they are attached. The chief effect of riding hydrogens is to make the structure factor calculation slightly more accurate by accounting for the small but measurable contribution of hydrogen towards the scattering of the crystal. If the hydrogen atoms are left out heavy atoms will move slightly in the direction of the absent hydrogens to account for the missing density. Riding hydrogen atoms become significant at around 1.5 Å resolution. A drop in

the R_{free} of around 1% is typical when riding hydrogens have been added (McRee 1999a). SHELX (Sheldrick 1998) is able to perform all these types of refinements and calculations making it ideal for high resolution refinement.

Radiation damage

The collection of high resolution data inevitability involves subjecting the protein crystal to large doses of radiation which has been shown to induce changes in the structure of the protein being determined. Breakage of disulphide bridges is the first known effect of radiation damage, which is preceded by de-carboxylation of aspartate and glutamate groups (Helliwell 1988, Ravelli and McSweeney 2000). Disulphide bridges are thought to be damaged first due to the large photoelectric absorption cross section of sulphur. Other structural effects of radiation damage include increased ADPs for damaged atoms and a possible swelling of the unit cell which is thought to be caused by the breakage of salt bridges. The atomic photoelectric absorption cross section (σ) can be calculated from the atomic scattering function f_2 .

$$\sigma=2r_0\lambda f_2$$

Where r_0 is the classical electron radius and λ is the wavelength in Å. If this calculation is performed for atoms likely to be in a protein crystal at a wavelength of around 1 Å, the results indicate that the photoelectric absorption cross section of sulphur is 80 times greater than that of carbon and nitrogen. During photoelectric absorption an X-ray photon is completely absorbed by a low level core electron, the which is then ejected from the atom carrying any excess energy as kinetic energy (Figure 3.17).

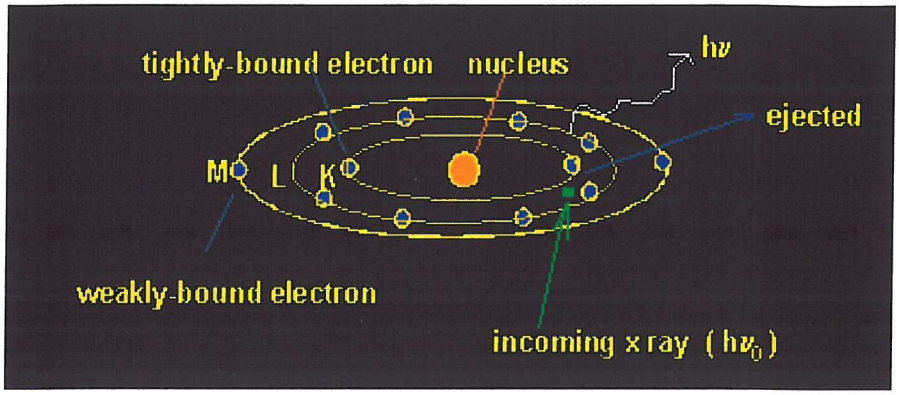


Figure 3.17 Illustrating the process of photoelectric absorption, the incoming X-ray is absorbed by a low level tightly bound electron which is then ejected from the atom with any excess energy being converted into kinetic energy. h is defined as Planck's constant while v is defined as the frequency.

The disulphide bridge in endothiapepsin is located on the outside of the protein and interaction with the solvent is possible thus damage to the disulphide bond by the free radicals formed by the destruction of a water molecules can occur. Disulphide bridges are strong electrophilic centres where an (RSSR-) radical is preferentially formed by the following reaction (Burmeister 2000) in which (p) represents the protein.

$(p)\text{-}CH_2SSCH_2\text{-}(p) + e^- \rightarrow (p)\text{-}CH_2SS^{*-}CH_2\text{-}(p)$

The (RSSR-) radical can be cleaved spontaneously by

$(p)\text{-}CH_2SS^{*-}CH_2\text{-}(p) \rightarrow (p)\text{-}CH_2S^- + (p)\text{-}CH_2S^*$

Or by a number of other mechanisms which involve attack of hydroxyl radicals.

$(p)\text{-}CH_2SSCH_2\text{-}(p) \rightarrow (p)\text{-}CH_2SS^{*+}CH_2\text{-}(p) + e^-$

$(p)\text{-}CH_2SSCH_2\text{-}(p) + OH^* \rightarrow (p)\text{-}CH_2SS^{*+}CH_2\text{-}(p) + OH^-$

As the disulphide bridge in endothiapepsin is exposed to the solvent it is not possible to say by which mechanism it may become damaged. However a water molecule belonging to the inner hydration shell is within 3 Å of the disulphide bridge. But the large photoelectric absorption cross section of sulphur makes the cleavage of the disulphide more likely to be caused by primary radiation damage, the main component of which is photoelectric absorption. Primary radiation damage is dependent only on the X-ray dose the crystal is exposed to, while secondary damage to protein crystals by free radicals is dose and time dependent. These free radicals can be generated by the destruction of water molecules.

The disappearance of any density between sulphur atoms has been observed before in high resolution time resolved structures for the protein Torpedo californica acetylcholinesterase (TcAChE) by Weik *et al* 2000. In this study following the disappearance of linking density between the sulphur atoms, one of the cysteine residues moves away from the other sulphur atom and after further radiation damage the sulphur detaches and finally disappears. While the other cysteine remains in place and its electron density gradually decreases.

Structure validation

In least squares refinement, as with all refinement methods, the contributions from X-ray data and geometrical data must be balanced. Too high a weight on X-ray data will give a reduced R_{factor} at the expense of geometry. Thus the R_{factor} of a model is only valid when accompanied by good geometry. This is usually defined by rms deviation from ideal values, for bond distances these should be less than 0.025 Å and 0.05 Å for bond angle distances. High B-factors for atoms within the hydrophobic core of the protein also indicate problems with the structure.

The progress of refinement is often monitored by the R_{factor}, which monitors the difference between the calculated structure factors (F_{calc}) generated from the current model and the observed structure factors (F_{obs}).

$$R = \frac{\Sigma(|F_{Obs}| - |F_{Calc}|)}{\Sigma|F_{Obs}|}$$

Throughout the refinement process there is a danger of over fitting the model, this may lower the R_{factor} but may not be meaningful. To avoid this a Free R factor (R_{free}) is used (Brünger 1992)

$$R_{free} = \frac{\Sigma(|F_{Obs}^{free}| - |F_{Calc}^{free}|)}{\Sigma|F_{Obs}^{free}|}$$

Usually 5-10 % of the observed structure factor amplitudes are chosen for inclusion in the R_{free} set either randomly or in resolution shells if non crystallographic symmetry (NCS) is present to avoid NCS related reflections being split between the refinement and R_{free} reflection sets. The reflections in the R_{free} are not used in the refinement process instead they are used only to calculate the R_{free} set and form the $|F_{obs}|$ in the above equation. The $|F_{calc}|$ values for data used in the R_{free} are calculated from the model refined with the remaining 90-95 % of refined structure factors. If the model is improving then both the R_{factor} and the R_{free} factor should drop roughly in parallel.

Chapter 4

Neutron diffraction theory and practice

The neutron (first discovered in 1912 by James Chadwick) is a neutral particle with spin ½ and a magnetic moment of 1.9132 nuclear magnetons. It has a mass similar to that of a proton and is stable within the nucleus of an atom. Outside of an atom a free neutron has a half life of around 17 minutes. When a neutron decays a proton, electron and an antineutrino are formed.

The energy and wavelength of a neutron are related to each other by the de Broglie equation.

$\lambda=h/(mv)$

Where λ is the wavelength, h is Planck's constant, m is the mass of a neutron and v its velocity.

Generation of Neutrons

There are two common ways to produce a beam of neutrons. One is Uranium^{235} based steady state nuclear fission, which produces a constant stream of neutrons some of which sustain fission and others are removed to produce a usable neutron beam. These excess neutrons are of a very high energy and are known as "hot or fast neutrons". They have a mean energy of ~1 MeV and are cooled or "thermalised" to 25meV via multiple collisions of the neutrons in a heavy water or graphite moderator kept at a steady temperature to produce neutrons with wavelengths between 1-3 Å. The heavy water moderator used at the ILL provides a large moderation length coupled with a low absorption cross section. This gives a broad neutron flux peak that occurs relatively distant from the core. The speed of the neutrons produced obeys a Maxwell distribution (a polychromatic or white spectrum) that is suitable for Laue diffraction.

Around 200 ~ MeV of heat are generated per fission event, this heat must be removed by an efficient cooling system. For reactor based system there is an inherent maximum neutron flux which is imposed by the reduced density of the

neutron generating material which can be effectively cooled and by the heat removal capacity of suitable coolants. There are a number of other factors which serve to reduce neutron flux such as the limited length and diameter of neutron guide tubes. These are limited to 3-4m and a diameter of 30cm respectively by radiation shielding and background reduction requirements.

The second common way to generate neutrons is using a "pulsed source" in which a cluster of charged protons from a linear accelerator are injected into a synchrotron and are condensed into a tighter pulse and then allowed to strike a metal target such as tungsten. The high-energy particles cause neutrons to be spalled or knocked out from the target nuclei in a nuclear process called spallation. Other neutrons boil off as the bombarded nucleus heats up. For every proton striking the nucleus, 20 to 30 neutrons are expelled (Schoenborn and Knott 2001). The neutrons from a spallation source are produced in accelerator based pulses which have a much higher intensity than that available from reactor sources. A number of spallation sources are currently planned or under construction such as the SNS at the Oak Ridge site in Tennessee that is under construction and is due to be finished in 2006. In the current design it is planned to use liquid mercury instead of tungsten in the target, which will allow more effective cooling to take place. A European spallation source (ESS) which would have a neutron flux around 6 to 10 times that of the ILL has also been proposed.

Neutron diffraction

Neutron diffraction studies can only be used with macromolecules if the structure of the molecule has been solved by X-ray crystallography. The main reason for this is the fact that the scattering length of neutrons is only variable by a small amount. This means that at present there are only limited methods to solve the phase problem with neutrons. This is because all current methods to solve the phase problem with macromolecules exploit large differences in scattering factor between different elements. The phases of the diffracted rays are needed to determine the positions of the atoms in real space and solve the proteins 3D

structure. This means that phases from earlier X-ray experiments must be used to provide phases for the initial calculation of neutron density maps. The simplest way to do this is to use the x, y, z co-ordinates of all atoms in the protein as a starting point for refinement with the neutron data. As hydrogen atoms have too few electrons to show up strongly in X-ray diffraction patterns their positions relative to other atoms are generated based on known bond lengths and angles. The sign of the neutron density at these generated hydrogen positions indicates whether the atom is hydrogen (negative neutron density) or deuterium (positive neutron density). A patch of positive density around a generated hydrogen atom would indicate that the hydrogen had exchanged for a deuterium. As neutrons carry no charge their velocity only arises from their temperature. Neutrons are diffracted only by the nucleus of an atom (for non magnetic atoms).

Incoherent Scattering

One major difference between X-ray and neutron diffraction relates to the change of phase of the reflected wavelet compared to the incoming wavelet. In X-ray diffraction this change is π (180°) and is the same for all atoms regardless of atomic number or isotope (Ladd and Palmer 1993). In neutron diffraction however this is not the case and different atoms and isotopes change the phase by either 0° or π (180°). A phase change of 0° is associated with a negative scattering length (incoherent scattering) while a phase change of π (180°) is associated with a positive scattering length (coherent scattering) as shown in Figure 4.00 (Ladd and Palmer 1993).

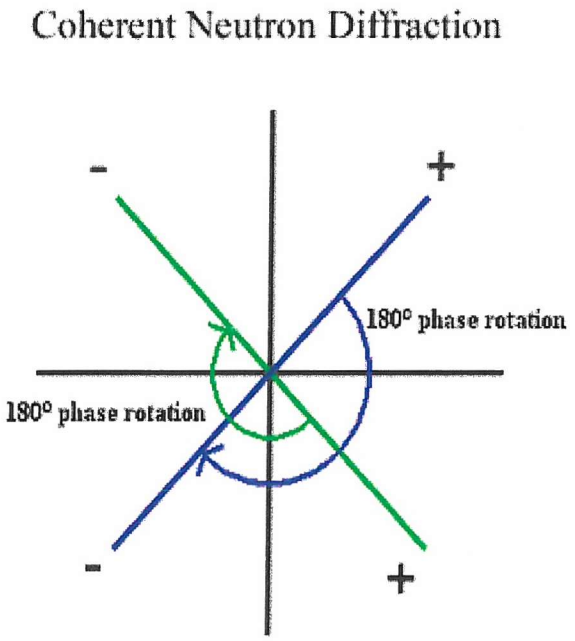


Figure 4.00 Showing the phase change associated with coherent neutron scattering 180° or π. There is no phase change associated with incoherent scattering between the incident and scattered neutron.

These differences in phase change induce destructive interference which is visible in the neutron density maps. Neutron density for methylene ($-CH_2-$) groups which posses two unexchangeable hydrogens is reduced as the coherent scattering from carbon is π out of phase with the incoherent scattering from the hydrogen atoms. For this reason there is frequently no density for methyl groups ($-CH_3-$) in neutron density maps with unexchangable hydrogens. Deuterium with its positive scattering length (+6.07 barns) can be observed more readily than hydrogen with its negative scattering length (-3.07 barns). As the positive scattering length of deuterium is twice the magnitude of the negative scattering length of hydrogen, deuterium density is more prominent in neutron density maps. In fact the overall scattering length is determined by two scattering lengths the coherent scattering length (b^+) and the incoherent scattering length (b^-) (Stuhrmann 2001). Both the incident neutron and the nucleus have a spin ½ . In the scattering of the neutron by the proton there are two different channels (shown in Figure 4.01).

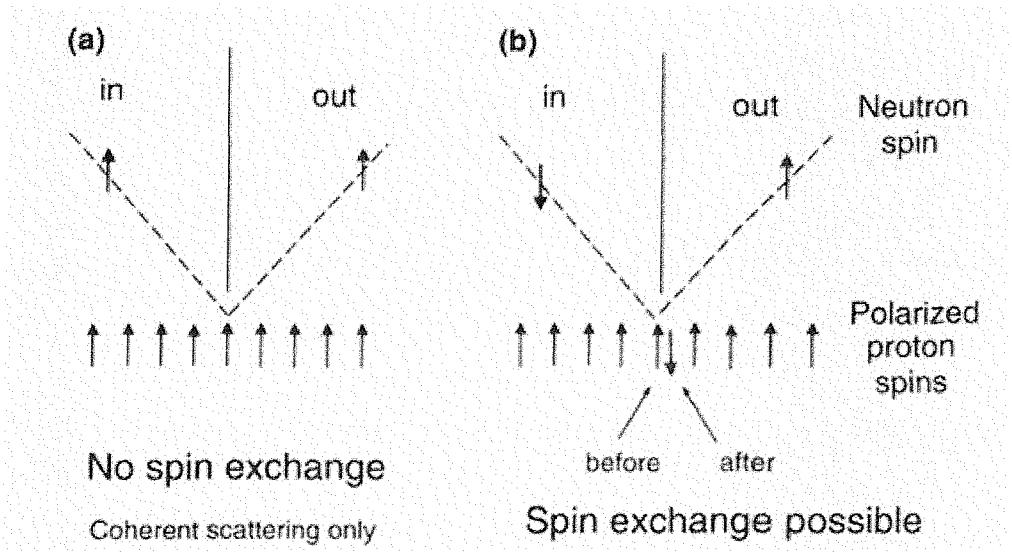


Figure 4.01 Scattering of a neutron by a proton, (a) The spin polarized protons spin in the same direction as the incident neutron, there is no spin flip (b^+). (b) The spin of the incident neutron points in the opposite direction to the proton spin, spin exchange is possible (b^-). (Taken from Stuhrmann 2000)

One for the total spin ½ + ½ =1 (b^+ scattering length) and one for the total spin ½ - ½ = 0 (scattering length b^-). The spin 1 has three substates 1,0 and -1 whereas there is only one state for the total spin 0. The effective neutron scattering length (b) of a proton is the weighted average of b^+ and b^-. With $b^+ = 1.083 \times 10^{-12}$ cm and $b^- = 4.74 * 10^{-12}$ for hydrogen we obtain

$$b^H = \tfrac{3}{4} b^+ + \tfrac{1}{4} b^- = -0.374 \times 10^{-12} \text{ cm}$$

This is the well known negative scattering length of hydrogen. Since most elements contain various isotopes each of which has a different neutron scattering length then a chemically pure sample in neutron diffraction behaves as if was made up of different species, resulting in diffuse elastic scattering. In neutron diffraction there is also a term related to the nuclear spin of the scatterer. Unless

nuclear spins are ordered (very difficult to achieve) and the neutrons spin polarized (halving the flux) incoherent elastic neutron scattering will always take place. These incoherent elastic scattering effects give rise to background intensity even for pure elements. Incoherent scattering does not contribute to reflection intensity but only to the background, for hydrogen atoms most of the scattering is incoherent. Only coherently scattered neutrons carry any information about the structure of the sample. The nucleus of an atom has a diameter in the region of 10^{-5} Å and as the wavelength of thermal neutrons is in the order of 1 Å there cannot be any interference caused by diffusion of the nucleus (which is often defined by a B-factor). In X-ray diffraction it is the diffuse electron cloud that is responsible for diffraction while the diffraction of neutrons arises from the much smaller nucleus. This difference relates to the atomic scattering factor, in X-ray diffraction as θ increases the atomic scattering factor lowers and the higher the atomic number of an atom the more pronounced this effect is. In neutron diffraction however a lowering of atomic scattering length (the neutron term for atomic scattering factor) is not observed at high θ values. The scattering length of an atom in neutron diffraction is also independent of atomic number. In X-ray diffraction at high θ values, atoms with high atomic numbers (Z) have greater path length differences between reflected wavelets associated with them. This increase in path length is associated with destructive interference resulting in a lower atomic scattering factor at high θ values. In neutron diffraction the path length differences between wavelets in atoms with high atomic numbers is very small due to the fact that the nucleus of at atom is much smaller than its electron cloud. This small size means that there is little difference in path length between wavelets diffracted from the same atom. Hence there is no drop off of the atomic scattering length at high θ values, also the scattering length is unaffected by atomic number. As only the nucleus takes part in neutron diffraction the amplitude of the diffraction depends mainly on resonance effects. These vary in unexpected ways from atom to atom and from isotope to isotope.

Thus the scattering lengths of neutrons can only be determined experimentally since there is no theory from which to predict them. The amount of absorption by matter also differs for neutrons compared to X-rays. With X-rays the rate of absorption increases dramatically as atomic number increases. In contrast with thermal neutrons the rate of absorption by matter is always much less but a slight increase in absorption does take place as atomic number increases. The fact that matter absorbs neutrons weakly gives rise to a number of important consequences. It permits the use of thick specimens up to a centimetre thick; this enables diffraction of large crystals to take place. This also helps to explain the fact that neutron diffraction causes little damage to the crystal enabling diffraction experiments to proceed for weeks at room temperature. While the damage to crystals caused by X-rays can be slowed with cryogenic cooling, crystal life span still remains well below that for neutron diffraction crystals.

Monochromatic neutron diffractometry is rarely performed with protein samples nowadays. One of the main reasons for this is that the diffraction is weak. The average intensity of a diffracted ray is proportional to the intensity of the incident radiation and to the ratio of the sample to unit cell volume. As neutrons are produced in relatively low fluxes and biological molecules are almost always small crystals with large unit cells, total diffracted intensity is weak and spread over a large number of reflections. This leads to a large distribution of weak signals (Myles *et al* 1998). In neutron diffactometry a large amount of background noise is present in the diffraction pattern and most of this comes from the sample itself, particularly any remaining hydrogen atoms. This leads to a low signal-to-noise ratio. Since a large proportion of noise in the diffraction images comes from hydrogen atoms in the sample itself, a high degree of exchange for deuterium is desirable. A number of different techniques are used to minimise the level of background noise in the diffraction pattern. The first one involves the use of large crystals whose volume is around $1 mm^3$. The second involves replacement of exchangeable hydrogen atoms in the protein crystal by soaking in D_2O or vapour diffusion. This leads to a positive signal from the hydrogen atoms, which have been replaced by deuterium significantly reducing the amount of background

noise. The large crystal size also increases the sample to unit cell ratio thereby increasing the intensity of the diffracted neutrons and increasing the signal-to-noise ratio (Myles *et al* 1998). The flux of neutrons from a thermal pile is irregular so to ensure the incident neutron beam is sufficiently intense wide beams must be used. These typically have a width of 3 millimetres; therefore the large crystals used in neutron diffraction are able to interact with most of the neutrons in the wide beam

Quasi Laue neutron diffraction

As the flux of neutrons is relatively low, the use of a monochromatic neutron beam would slow down the experiment further. Thus a polychromatic neutron beam can be used in an experiment to lower the time needed to collect a complete data set by increasing the flux of neutrons on the sample. Typically an 8-20 % wavelength section is used from the white beam (Nimura *et al* 1997). The major disadvantage of this technique is the production of multiplets (superimposed spots). However as only an 8-20 % section of the beam is used this reduces the number of multiplets to less than one percent of the observed reflections. This means that most of the low-resolution reflections are preserved for analysis. In normal Laue methods where a wider range of wavelengths are used, a significant number of the low resolution reflections cannot be analysed, as they are present as multiplets. Also the use of 8-20 % of the white beam cuts down the amount of background noise. If the whole white beam were used, the background noise would be at a high level and there would be extensive overlap between reflections. As only 8-20 % percent of the white beam is used this method is referred to as quasi Laue. Further benefits for proteins with larger unit cells when using just 8-20 % of the beam include fewer spatially overlapping reflections. The reduced incident spectrum produces a reduced number of reflections per image, thus cutting down on the number of reflections lost to spatial overlap. Thus the low number of spatial multiplets helps to increase the completeness of the data.

LADI detector

The LADI (LAue DIffractometer) detector as shown in Figure 4.02 is a cylindrical large area neutron detector with the image plate being formed from four smaller image plates. The detector is an integrating device well suited to data acquisition over long periods. The neutron image plates (NIPs) are formed when a neutron converter such as Gd_2O_3 is combined with a photostimulated luminescence material on a flexible plastic support (Wilkinson *et al* 1992).

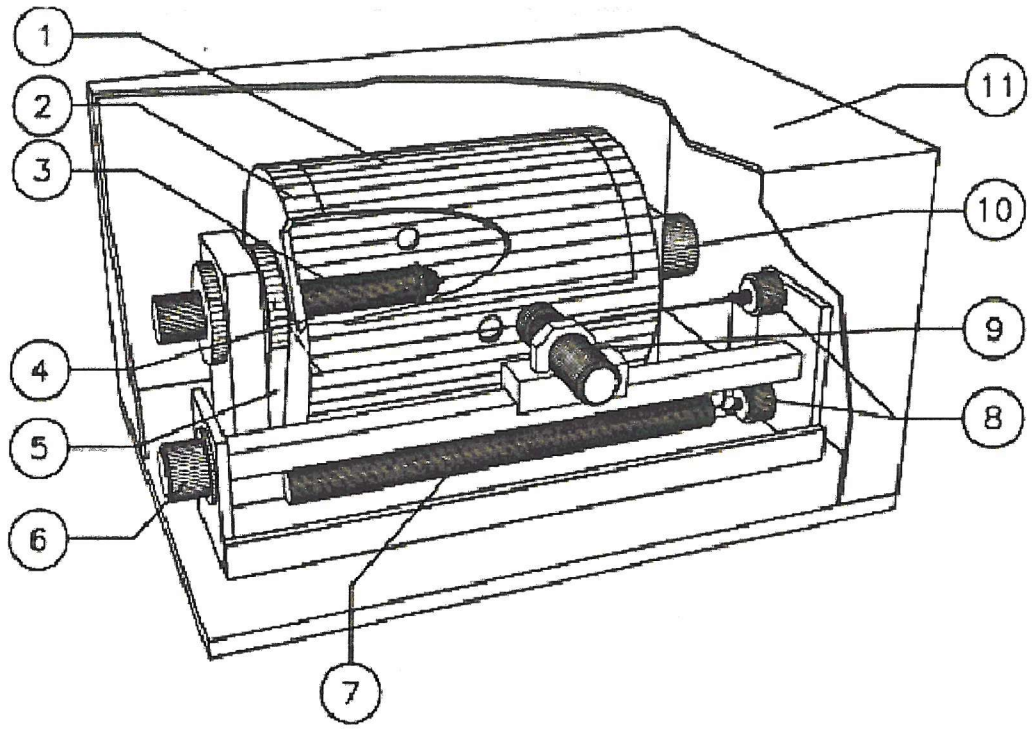


Figure 4.02 The LADI detector. 1: Image plate on drum. 2: Drum. 3: Sample holder. 4: Crystal. 5: Transmission belt to drive drum. Motor is under table. 6: Carrier for reading head with photomultiplier. 7: He-Ne laser. 8: Mirrors for bringing the laser light to the reader head. 9: Reader head with photomultiplier. 10: Encoder for drum rotation. 11: Cover. (Taken from ILL homepage www.ill.fr)

The neutron image plate is made up from four 20cm x 40cm Gd_2O_3 doped image plates. An example of a Laue diffraction image recorded on these plates is given in Figure 4.03.

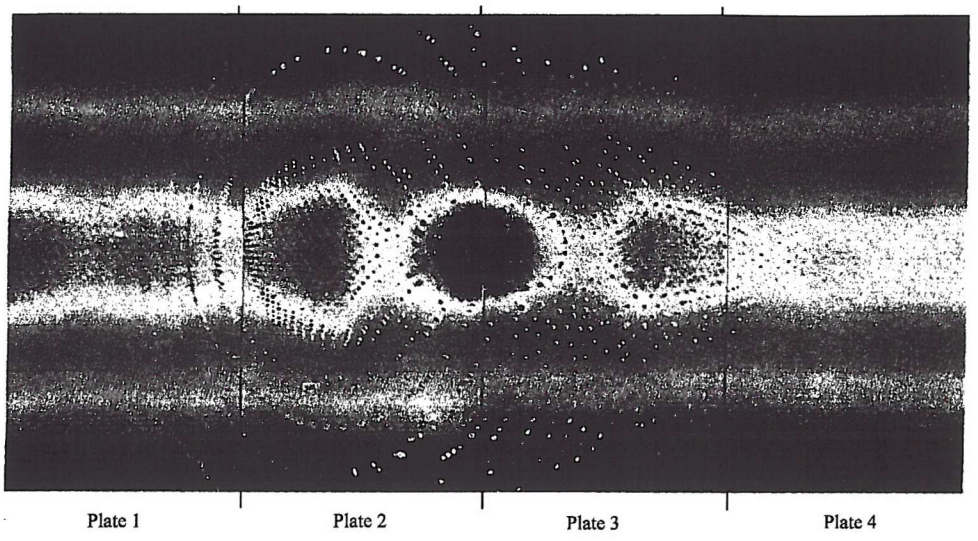


Figure 4.03 An example of a Laue diffraction pattern recorded on four neutron sensitised image plates.

The NIPs are placed around the outside of an aluminium cylinder which has a diameter of 31.8cm and a length of 40cm. A central hole is present in the cylinder, this allows the incident neutron beam to enter the detector and interact with the crystal, which is placed within the cylinder. The undiffracted neutrons from the incident beam pass out of a hole in the cylinder directly opposite to the hole for the incoming incident beam. The four NIPs cover 80 % of the outside of the aluminium cylinder. Neutrons are able to pass straight through aluminium and thus pass through the cylinder where they interact with the NIPs. The cylindrical design of the LADI detector enables almost all reflections from the crystal to be detected. Consequently there are very few unrecorded reflections; this helps to reduce the time taken to collect a complete neutron data set. A complete data set can be collected from only a few orientations of the crystal. The neutron beam used in LADI experiments comes from a heavy water moderated uranium reactor on site at the ILL. For studies of proteins a narrow wavelength spectrum of neutrons is used.

This is achieved using a Ti/Ni multilayer bandpass filter, which sets the neutron wavelengths used in the experiment between 2.60 and 3.60Å.

Experimental procedure

The width of the beam used in this experiment was controlled by the use of a collimator placed in the incident beam. The diameter of collimator used in this experiment was 2.9 mm. The crystal used was wet-mounted in a 2 mm wide capillary tube with a small reservoir of mother liquor in the capillary on each side of the crystal. The ends of the capillary where sealed with dental wax. A total of 30 images had been collected previously on this crystal. The purpose of this experiment was to collect more diffraction data in order to increase data completeness, which at the start of the experiment was at 79.9 % at 2.10 Å. The endothiapepsin H261 co-crystal was mounted and centred on a goniometer whose head was in line with the cylinder axis of the LADI detector. The crystal was rotated 15° around this axis between each exposure; no rotation of the crystal took place during exposures. The H261 co-crystal was mounted within the LADI drum and the crystal exposed to the neutron beam for a period of 30 hours. After each exposure the neutron beam was halted, and a He/Ne laser and read head were used to read off the diffraction image present on the NIPs in a phonographic manner. This proceeds with the head slowly tracking horizontally while the cylinder is rotated at high speed. The readout time is five minutes, giving an image of the pattern recorded on the plates comprising of 4000 x 2000 square pixels 200μ on edge. Since the Bragg peaks are read out from the rear surface of the NIP on which there is a polymer support film, this decreases the effectiveness with which the Bragg reflections are recorded. The main reason for this is that the readout laser must read the NIP from the back, this increases the penetrating length which is associated with a decrease in the intensity of each Bragg reflection. The intensity of a Bragg reflection is reduced the further away from the internal surface it is recorded.

Following the scanning the NIPs are erased by the application of intense light from a bulb present within the LADI detector in the same phonographic manner. The crystal axis is then rotated around by 15° then neutrons are then allowed back onto the crystal. Endothiapepsin with its relatively small unit cell dimensions is ideal for exposure on this type of detector. Diffraction from proteins with a larger unit cell could increase the spatial overlap of reflections. The level of spatial overlaps is very important because of the cylindrical detector shape, which means the crystal to detector distance is fixed. A unit cell of large size would give more reflections giving increased spatial overlap of reflections. In the case of the LADI detector, background noise on the image plate can be caused by any gamma rays present when the experiment is taking place. Work is underway at present to make the LADI detector less sensitive to external gamma rays. The LADI detector is located some distance from the reactor core; this serves to reduce background from the reactor.

Processing of the Neutron Laue diffraction images

In Bragg's law the wavelength of the diffracted beam is required to calculate the length of the reciprocal lattice scattering vector (**S**). For a Laue experiment this value is not known, this means that a different strategy is required to work out cell parameters in a Laue experiment compared to a monochromatic diffraction. If however the cell parameters are known e.g. as they were for the endothiapepsin H261 co-crystal, they can be entered directly into LAUEGEN (Campbell *et al* 1998). The LAUEGEN software is used to index the Laue diffraction patterns. The next stage of data processing is to find the orientation of the unit cell in the neutron beam.

The program was able to find the unit cell orientation via auto-indexing. In order to do this the position of three or more nodal reflections where entered by eye (Figure 4.04). A nodal reflection is a super-position of many reflections from common reciprocal lattice points of two or more centric zones.

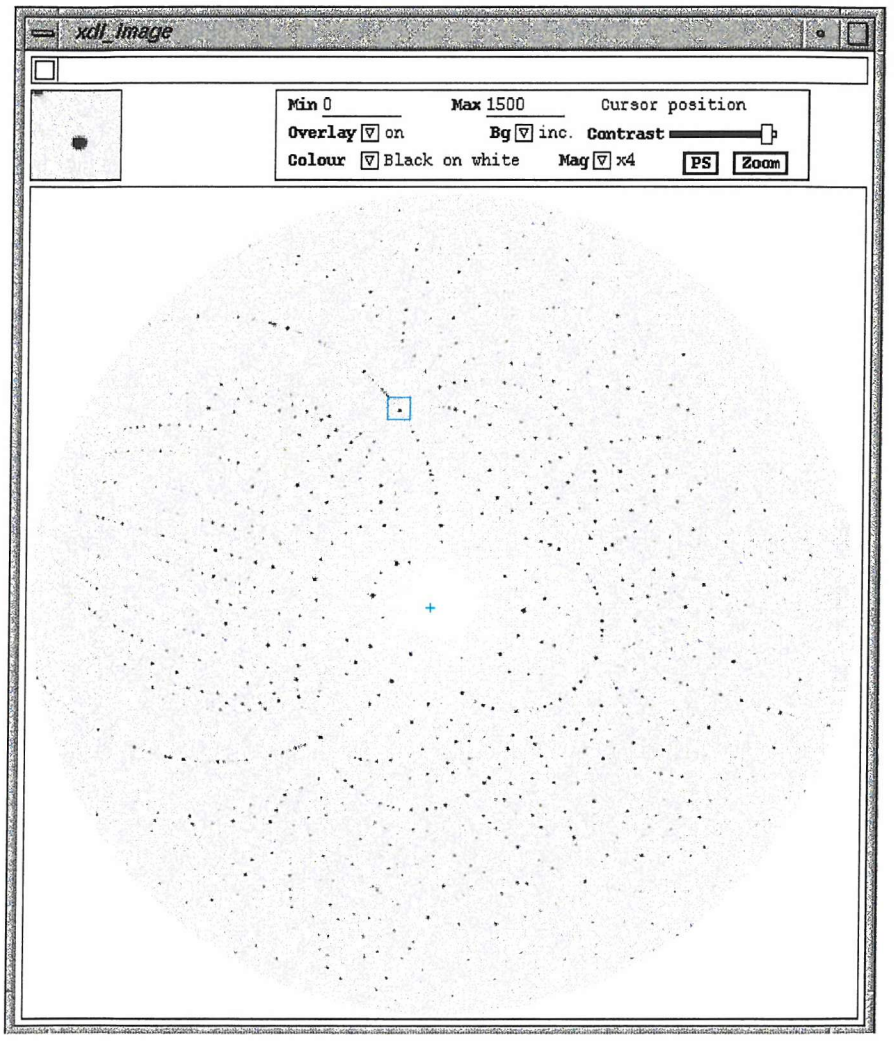


Figure 4.04 A typical X-ray Laue diffraction image, determination of unit cell orientation relies upon the identification of nodal reflections similar to the boxed reflection. Taken from the Lauegen homepage.

After the selection of nodals on the diffraction pattern a number of possible solutions where produced by LAUEGEN. The correct solution that predicted nearly all of the visible reflections was selected. This solution was the starting point for refinement of the unit cell orientation. The unit cell orientation as well as the plate characteristics are then recorded in a Laue Data Module (LDM) file. This is a text file, which contains values for the refined crystallographic parameters used to index the diffraction pattern. One of these files can be produced for each of the images recorded. Following this the diffraction pattern can be processed further in LAUEGEN by determining the average spot size, the orientation of the unit cell and other parameters being refined using the positions of nodal reflections.

In the first round of refinement the image plate parameters are refined. In the next series of refinements the unit cell dimensions are allowed to refine with the image plate parameters. In the third series of refinements the distortion corrections are also added to the refinement. All these refinements seek to minimise the RMSD in mm between the predicted positions of nodal reflections and their actual positions. After each refinement cycle the current rms and starting rms values are given. This indicates whether the refinement had succeeded in lowering the rms value and whether a further refinement round is required. After this section is complete and the rms reflection deviation values for the nodal reflections have been reduced to a minimum, the soft limits for the diffraction pattern are refined. The soft limits refinement seeks to improve the values for the minimum wavelength (λ_{min}) and the diffraction limit (d_{min}) by measuring integrated intensities for an over predicted pattern and analysing the resultant intensities as a function of the wavelength or resolution. This gives improved values (λ_{min}) and (d_{min}).

The program LAUESCALE was then used to normalise the raw integrated Laue intensity data to yield fully corrected structure amplitudes for both data sets using a wavelength normalisation curve. The wavelength normalisation curve was derived from internal Laue data using symmetry equivalent reflections recorded at different wavelengths present in the input mtz files.

Chapter 5

Results of the Neutron and X-ray diffraction experiments

Neutron diffraction of Endothiapepsin H261 complex

For diffraction images 1 to 6 of the 12 image data set collected in April 2000 the following optimal values were obtained for the soft limits.

d_{min} 2.20

λ_{min} 2.64

λ_{max} 3.60

Following this the spot intensities where integrated and intensity files written out as unsorted unmerged mtz files. After the initial processing of the first six diffraction patterns from the LADI by LAUESCALE and inspection of the wavelength normalisation curve (Figure 5.00) it became clear that neutrons with wavelengths of around 3.14 Å were not reaching the LADI.

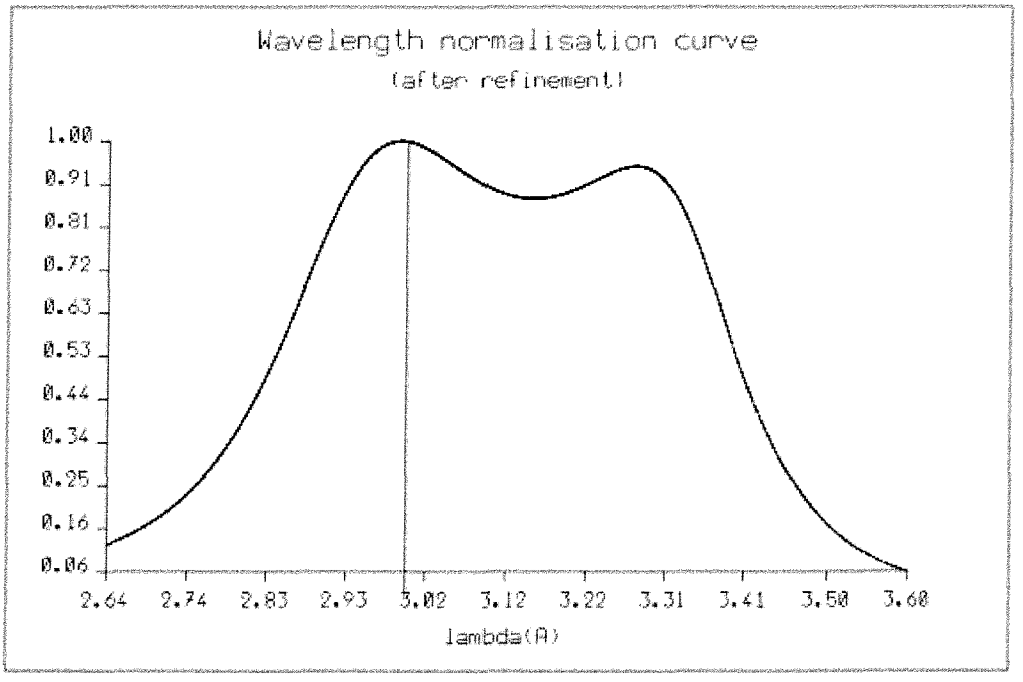


Figure 5.00 The wavelength normalisation curve for the first six images collected from the LADI. There is an obvious dip in the number of neutrons with wavelengths of around 3.14 Å.

This was attributed to an inelastic scattering machine further up on the neutron beam line, which had been using neutrons of around 3.14 Å wavelength. This meant the number of coefficients used to calculate the wavelength normalisation graph had to be increased from 6 to 40 to better model the wavelength profile of the incoming neutron beam (Figure 5.01). This was achieved by altering the parameter file for the program LAUESCALE.

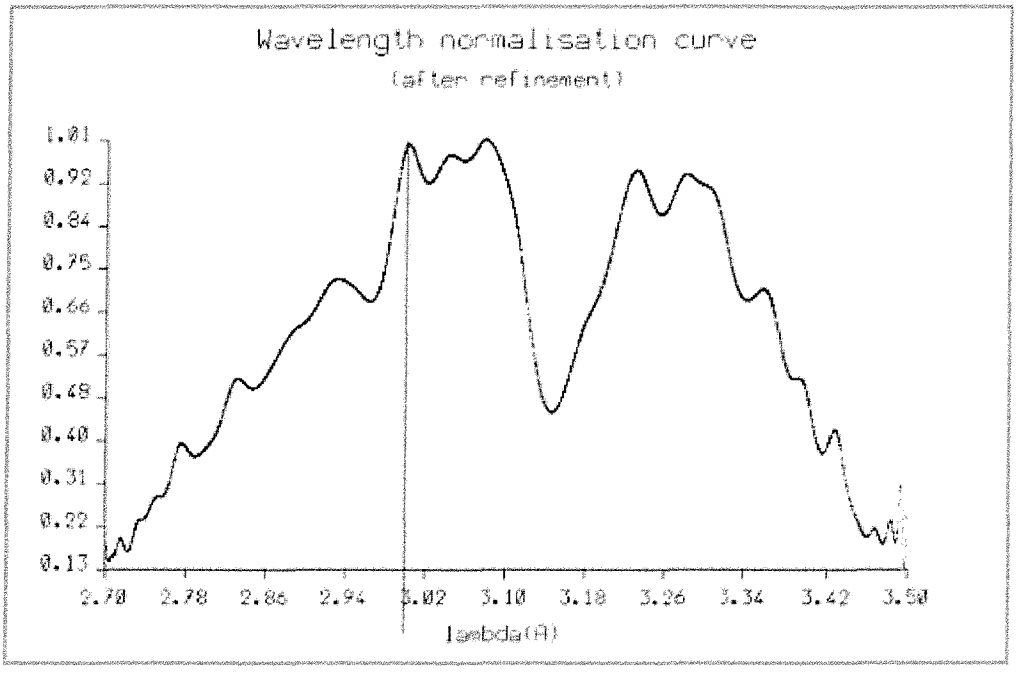


Figure 5.01 Showing the effect of increasing the number of coefficients from 6 to 40 to better model the wavelength profile of the incoming neutron beam.

The calculated average d_{min} value for images 1 to 6 was 2.20. These six images were processed separately from images 7 to 12 because of the differences in the wavelength profile of the incoming beam.

Images 7 to 12 where recorded with the full wavelength spectrum of neutrons from 2.64 to 3.60 Å producing six images with the following average values

d_{min} 2.10

λ_{min} 2.64

λ_{max} 3.60

As can be seen there is a reduction of 0.10 Å in d_{min} attributable to the increased neutron flux for the second data set (Figure 5.02).

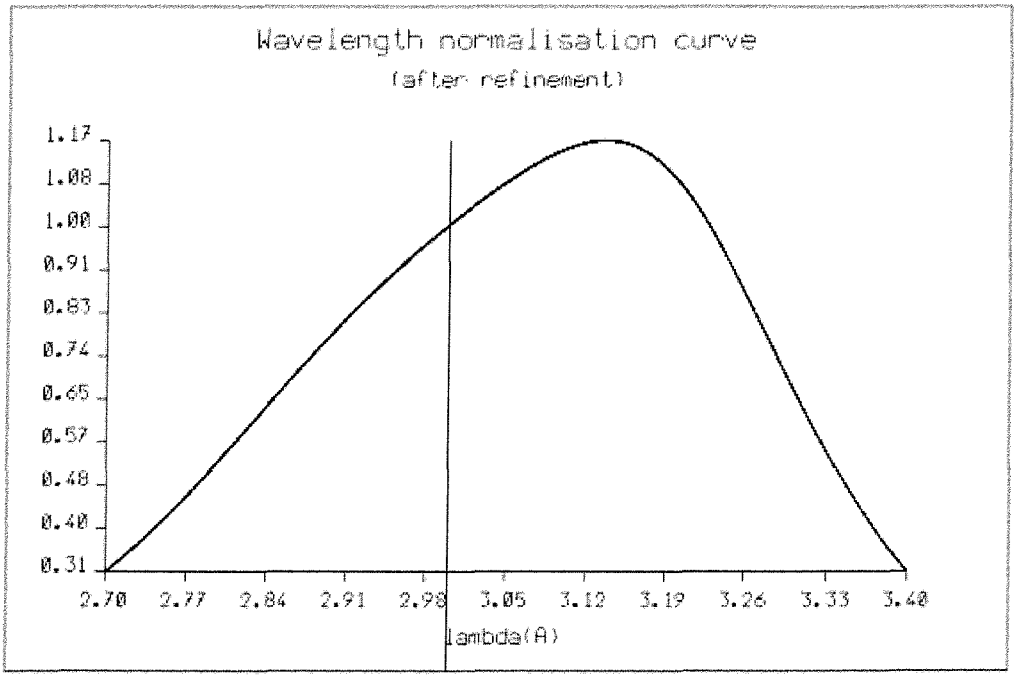


Figure 5.02 The wavelength normalisation curve for images 7-12 which shows the expected wavelength profile.

The parameter file for LAUESCALE was set so as the wavelength normalisation curve would be calculated between 2.64 to 3.60 Å, and deconvolution of multiple reflections was turned off. Spatially overlapping reflections were also not processed and an $I/\sigma(I)$ rejection cut off of 2 was used for output reflections. This was done for both data sets and helped to smooth out the wavelength

normalisation curve for the first data set. The unmerged reflection files for images 1 to 6 and 7 to 12 were output into two separate mtz files by LAUESCALE and these were then merged into an existing mtz file containing reflections from previous data sets using the merge command of MTZUTILS. In order that the three mtz files could be merged, the batch numbers of the first data set containing data from images 1 to 6 where increased by 300 using the CCP4 program REBATCH. The batch numbers in the second mtz file formed from images 7 to 12 where increased by 400. The reflections in the merged mtz file were then sorted by their h, k, l indices using the CCP4 program SORTMTZ. Scaling of the new data sets against existing data on the same crystal took place using the CCP4 v4.01 program SCALA. After scaling an mtz reflection file was produced in which 5 % of the reflections were given an R_{free} flag using the FREERFLAG program from the CCP4 suite (CCP4 1994). These reflections did not under go any further refinement and were used to calculate an R_{free}. This mtz file was then converted into ASCII hkl format suitable for refinement within SHELX by the CCP4 program MTZ2VARIOUS. This file formed the base for further work involving a modified version of SHELXL 97-1 (Sheldrick 1998), which had been altered to work with deuterium and hydrogen atoms. SHELX requires two input files: the .hkl reflection file that contains the h, k, l intensity (I) and σ(I) values for each reflection. And an .ins file which contains a series of restraints and constraints for the bond lengths and bond angles between atoms in various amino acid residues and other non standard molecules found in the inhibitor. The .ins file also contains the atomic co-ordinates of all the atoms in the molecule, the unit cell dimensions and the space group of the crystal as well as the instructions for the refinement protocol.

After a SHELX refinement is run, a number of output files are produced. One of these is the pdb file that can be loaded into any protein viewer program. The pdb viewer used was the 1996 edition of TURBO-FRODO (Bio-Graphics, Marseille). While this displays the structure of the protein, the assignment of which hydrogen atoms have exchanged with deuterium atoms must be done by the manual inspection of DF_o-mF_c neutron density maps. The program SHELXPRO which is

able to calculate TURBO-FRODO maps can produce these from the .fcf (reflection data) and .pdb files which are produced by SHELX during the refinement process. Two Fourier maps were produced using SHELXPRO; a $2mF_o$-DF_c map and a mF_o-DF_c difference map. These maps were loaded into TURBO-FRODO. The mF_o-DF_c map was contoured at -2.5 σ and +2.5 σ while the $2mF_o$-DF_c map was contoured at +1.2 σ. All exchangeable hydrogen atoms (hydrogen atoms linked to oxygen or nitrogen atoms) were checked over to see if a patch of positive neutron density from the mF_o-DF_c map was found in a suitable position. If so then the hydrogen was replaced with a deuterium atom. The position of the starting hydrogen atoms in the model were generated by SHELX using known bond lengths and angles.

The final neutron refinement R_{factor} is 23.5 % and 27.4 % for the R_{free}. These R-values are slightly high in comparison with the values expected for a refined X-ray structure at comparable resolution. However, they are consistent with R-values obtained in other neutron analyses. It should be remembered that a neutron diffraction structure has twice the number atoms found in an X-ray structure due to the addition of hydrogen atoms. The data processing and refinement statistics are shown in Table 3.

PDB code	1gkt
Unit Cell	a 43.1 Å, b 75.7 Å, c 42.9 Å, β 97.0°
Space Group	$P2_1$
Number of unique reflections	13,548
Resolution Range Å	20-2.10(2.2-2.1)
Multiplicity	3.4(2.4)
Mean I/σ(I)	5.4(1.7)
R_{merge}	7.5 % (11.9%)
Data completeness	84.5 % (72.6 %)
Number of reflections	13,548
Overall R_{factor} (%)	23.46
Overall R_{free} (%)	27.42
RMSD bond lengths (Å)	0.009
RMSD bond angles (Å)	0.014
RMSD bumps(Å)	0.051
RMSD chiral volumes ($Å^3$)	0.027
RMSD planes (Å)	0.023
Number of atoms in asymmetric unit.	
C	1569
N	379
O	772
S	2
H	2055
D	393
Total	5170
Data to parameter ratio	1.19

Table 3 Crystallographic statistics for the endothiapepsin H261 structure. Figures for the outer shell are given in brackets.

A Ramachandran plot for the neutron structure is shown in Figure 5.03. The neutron structure superimposes very well with the previously solved X-ray structure of this complex (Veerapandian *et al* 1990); the rms deviation between the two is only 0.2 Å for all α C atoms.

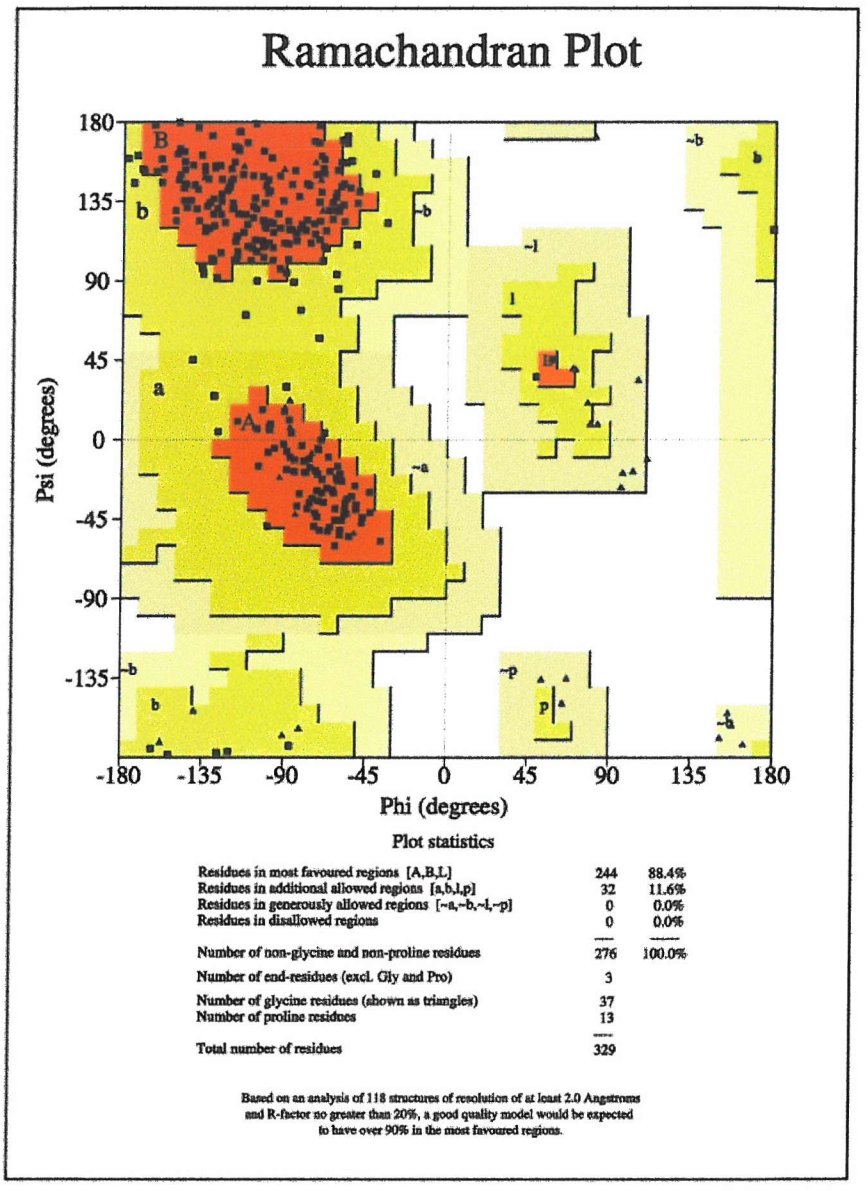


Figure 5.03 A Ramachandran plot of the endothiapepsin/H261 structure produced from the final refinement of the neutron data

The majority of amino acids have become deuterated either in the side chain or in the main chain with the exposed secondary structure elements being the most affected. In total 215 residues out of 330 (65 %) have undergone side chain and/or main chain deuteration. In the backbone, a total of 161 main chain amides (49 %) have exchanged. The parts of the molecule most protected from exchange are the buried β-strand regions. In contrast, the strands which have become most deuterated are generally those at the exposed edges of the sheets. The majority of polar side chains have exchanged and in general the loops and helical regions have exchanged to a greater extent than the β-sheet regions. Although the active site cleft is occupied by a tight-binding inhibitor, many residues in the vicinity of the catalytic centre and specificity pockets of the enzyme have undergone H-D exchange (Figure 5.04). It is perhaps surprising that the protein has not become deuterated to a greater extent. However, it should be remembered that the crystals have a low solvent content (39 %) and the efficiency of capillary vapour diffusion for H_2O/D_2O exchange is not well characterized as yet (Bon *et al* 1999).

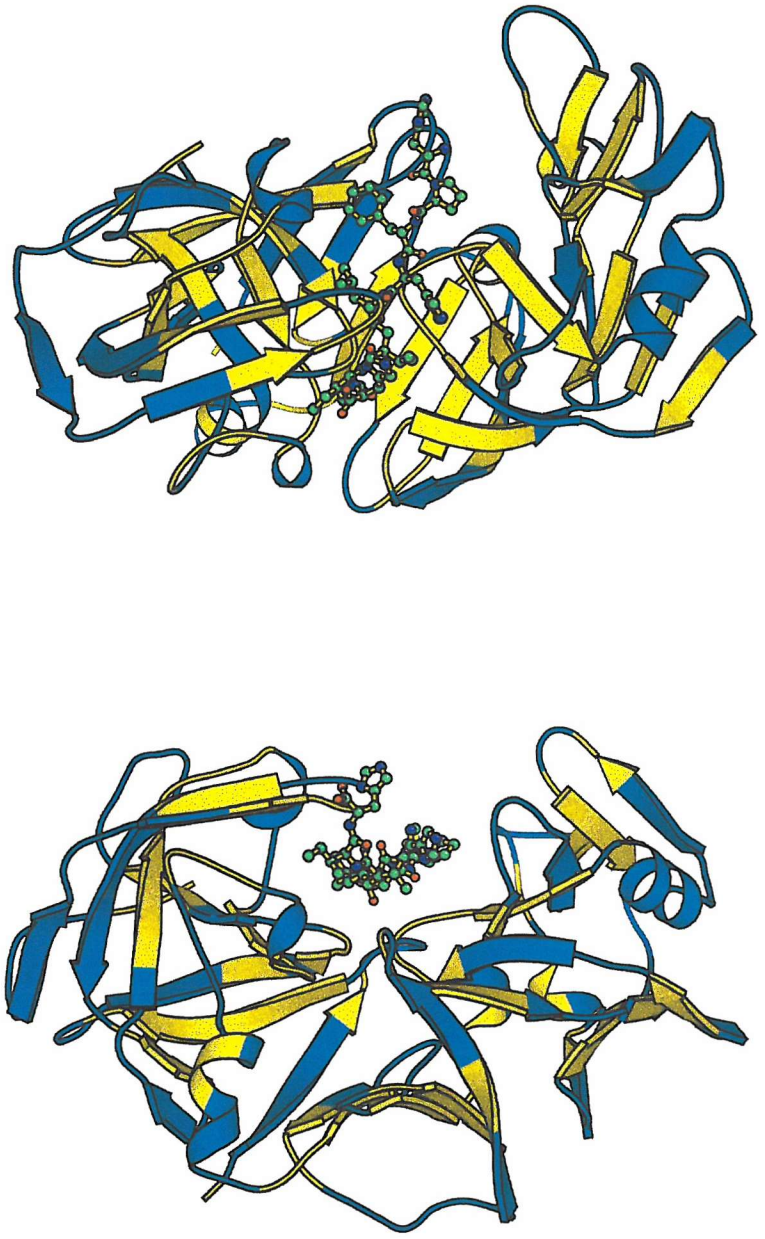


Figure 5.04 Showing the extent of main chain deuteration of the endothiapepsin molecule in two orthogonal views. Yellow shows the regions that have not exchanged whereas those shown in blue are segments where the amino acids have become deuterated in the main chain. In each view the inhibitor (H261) can be seen occupying the active site cleft.

Water structure

The structure contains 4674 protein atoms (including hydrogens and deuteriums), 159 inhibitor atoms and 256 solvent sites; 42 of these solvent sites have well-defined density for D_2O molecules and were therefore modeled as D_2O (Figure 5.04a). Whilst it might be expected that more D_2O sites would have been apparent at this resolution, the majority of solvent density peaks were spherical and could be refined satisfactorily as O atoms as is usual in X-ray studies (Figure 5.04b) . The neutron visibility of the deuterium atoms in solvent molecules would be reduced if they have high temperature factors. H_2O molecules scatter neutrons very weakly due to cancellation of the H and O scattering lengths which have opposite sign. Thus the solvent sites which appear to only have density for the central O atom are likely to be orientationally disordered D_2O molecules. The majority of the well-defined D_2O molecules are close to the surface of the protein and some are partially buried within the protein.

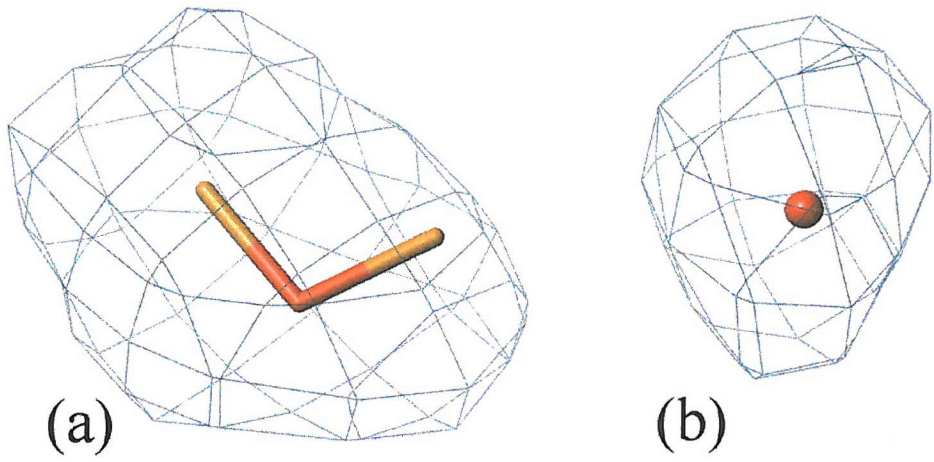


Figure 5.04 (a) The +1.2 σ $2mF_o$-F_c density for typical D_2O molecule (b) the 1.2 σ $2mF_o$-F_c density for a water molecule. The density associated with the D_2O molecule is more elongated than the sphere-like density features that were modeled as water molecules.

Of the 42 D_2O molecules modeled into the structure most of them (29) are within the inner hydration layer; 18 of the D_2O sites in the neutron structure have been modeled as water molecules in a room temperature 1.6 Å X-ray structure of endothiapepsin complexed with H261 (Veerapandian *et al* 1990). These sites are also conserved in a cryo-cooled 130 K 1.1 Å structure of endothiapepsin bound to H261. Of the 233 water molecules modeled into the neutron structure only 79 have a corresponding molecule in the room temperature 1.6 Å X-ray structure of H261 bound to endothiapepsin. This low level of agreement in which only a third (33.91 %) of the neutron water sites could reflect the mobility of the bulk solvent within the crystal at room temperature. The water structure of the room temperature 1.6 Å H261 structure was compared against the water structure in a cryo-cooled 1.1 Å H261 structure (Erskine *et al* unpublished). Of the 322 water molecules in the room temperature structure, 245 (76.1 %) have a corresponding water molecule in the 1.1 Å structure with most of these water molecules being outside the inner hydration shell. Differences between water molecules within the inner hydration shell could often be attributed to differences in the orientation of hydrophilic sidechains on the surface of the protein.

Deuteration of the active site residues

The residues forming the catalytic centre of aspartic proteinases are two strongly conserved Asp-Thr-Gly-Ser/Thr sequences. These are provided by the two domains of the enzyme where they associate to form the active site cleft. The two aspartate carboxyls are involved in numerous hydrogen bonds which keep them approximately co-planar. The outer oxygens of the aspartate diad accept hydrogen bonds from the Ser/Thr side chains in the above consensus sequence. It is clear from inspection of the neutron maps that Ser 35 and Thr 218 residues have deuterated side chains in spite of having very low solvent accessibility due to the presence of the inhibitor in the active site. The refined neutron structure and maps for the active site are shown in Figure 5.05.

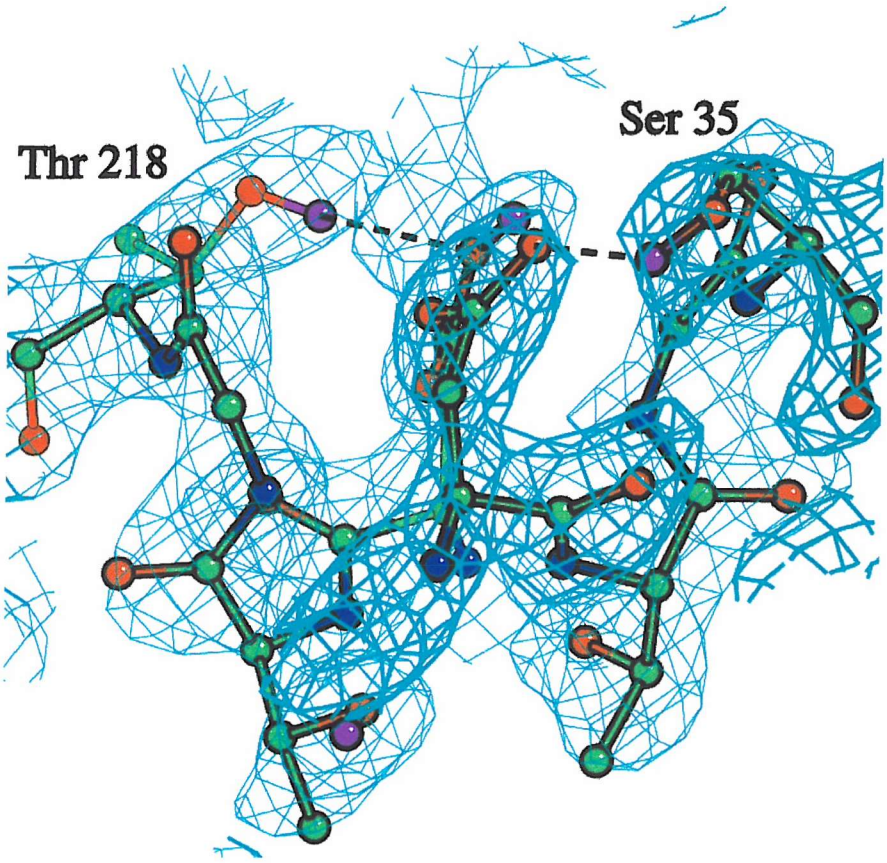


Figure 5.05 Showing the extent of deuteration of residues at the active site. The $2mF_o$-DF_c density map is contoured at +1.2 σ.

Since the active site flap and the inhibitor help to shield the catalytic residues from the surrounding solvent, the fact that buried active site residues have become deuterated may indicate that the network of hydrogen bonds in the active site cleft provides a means of exchanging protons with the bulk solvent. The σ_A weighed $2mF_o$-DF_c density at 1.2 σ around the active site is shown in Figure 5.06 for an unbiased model with no protons in the active site coloured light blue. The σ_A weighed mF_o-DF_c density is also shown at 2.5 σ (dark blue) and -2.5 σ (red).

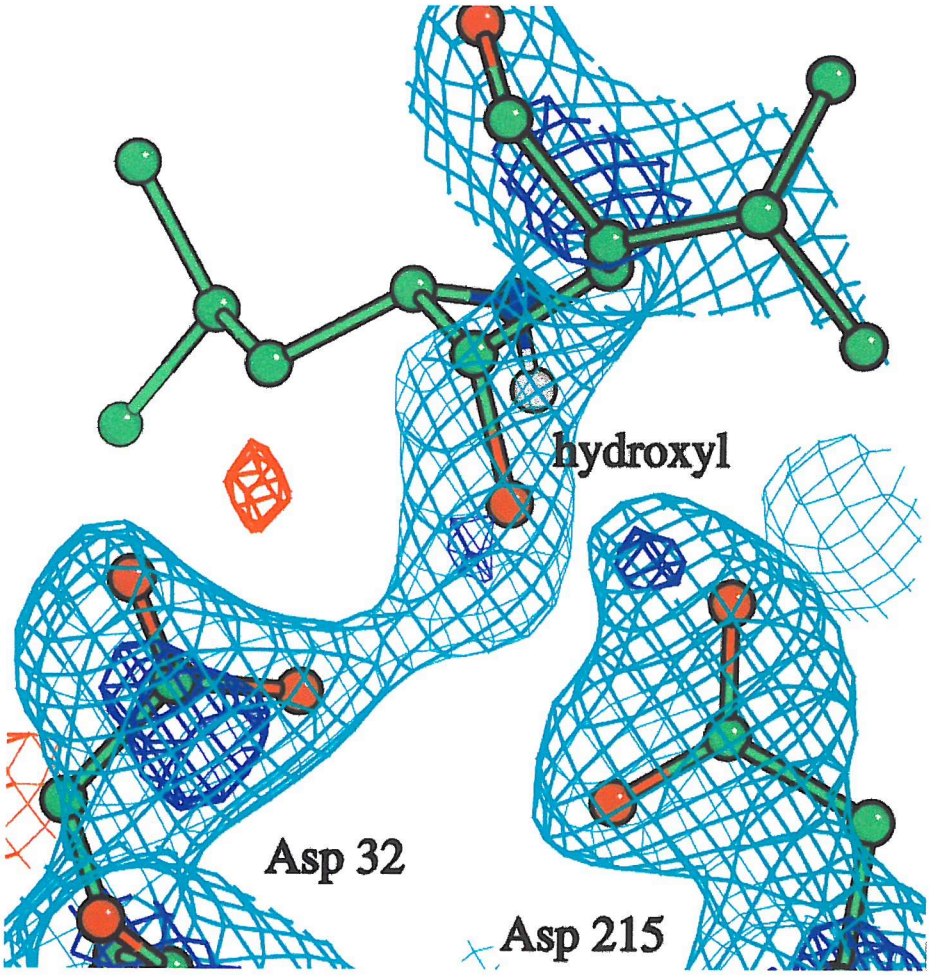


Figure 5.06 An unbiased view of the σ_A weighted $2mF_o$-DF_c at 1.2 σ in cyan and the mF_o-DF_c density at +/- 2.5 σ in the active site in blue and red respectively. There is positive density suggesting the deuteration of Asp 215 on its outer oxygen and for the location of the deuteron on the statine hydroxyl.

Two models were then generated to test the proton positions on the active site residues one with Asp 215 protonated on its outer oxygen and Asp 32 negatively charged and one with Asp 215 negatively charged and Asp 32 protonated with a deuterium atom on its inner oxygen. The resulting density around the active site for these two models is shown in Figures 5.07 and 5.08. With the map colouring scheme being the same as for model 1 the σ_A weighted $2mF_o$-DF_c density at 1.2 σ around the active site is coloured light blue. The σ_A weighed mF_o-DF_c density is also shown at 2.5 σ (dark blue) and -2.5 σ (red).

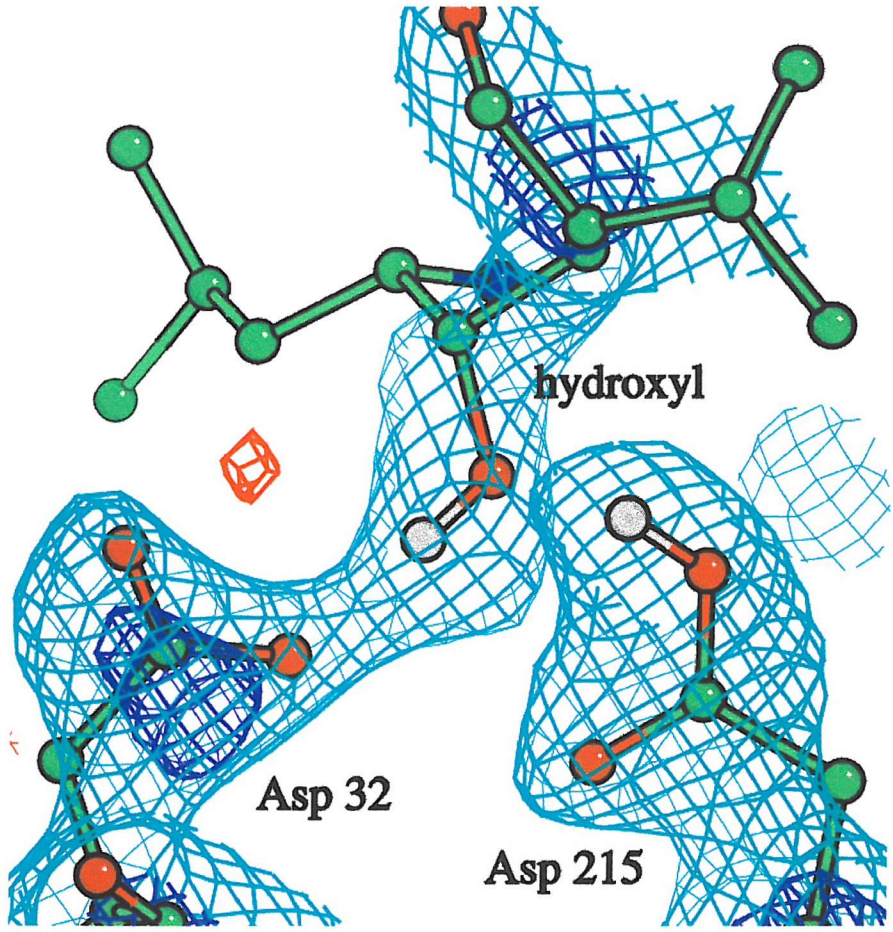


Figure 5.07 Showing the active site modelled a deuterium on the outer oxygen of Asp 215. With the $2mF_o$-DF_c density contoured at 1.2 σ shown in cyan and the mF_o-DF_c density contoured at +/- 2.5 σ shown in blue and red respectively. This model explains the two positive patches of density found in the unbiased model.

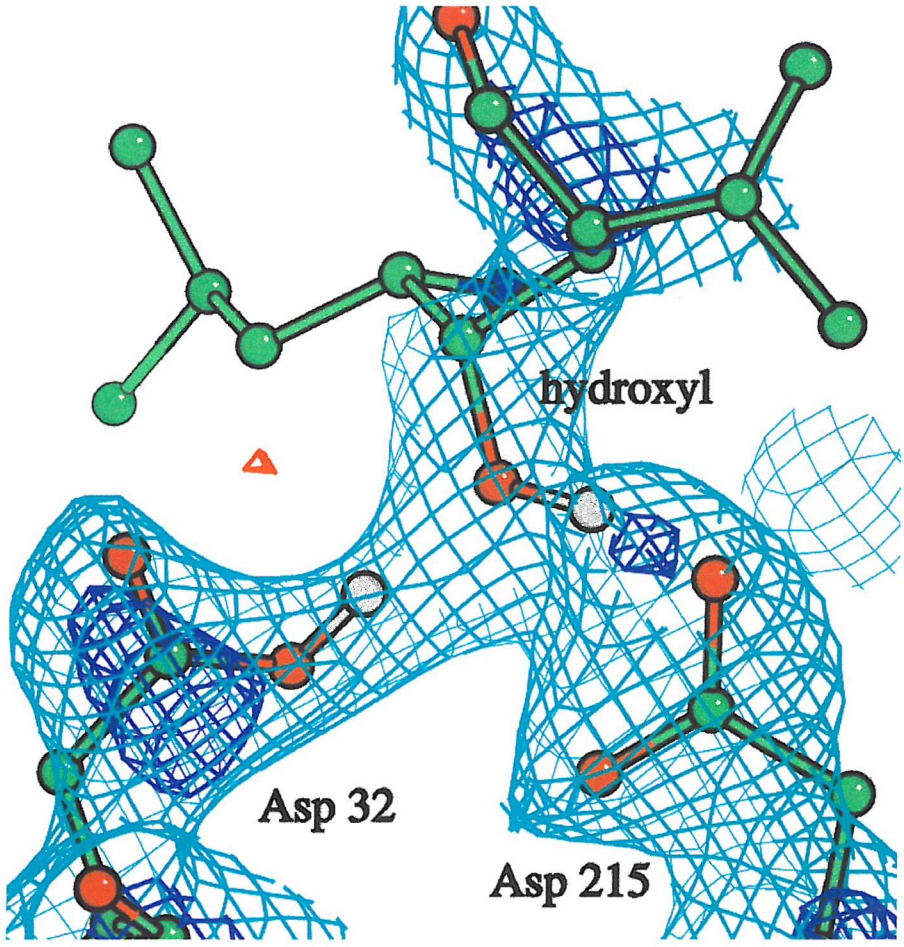


Figure 5.08 Showing the active site modelled with a deuterium proton on Asp 32. With the $2mF_o$-DF_c density contoured at 1.2 σ shown in cyan and the mF_o-DF_c density contoured at +/- 2.5 σ shown in blue and red respectively. The small patch of positive density close to the outer oxygen of Asp 215 also present in the unbiased model is not explained.

The occupancy for the protons in each of the two models was refined using SHELX, the results of this occupancy refinement are shown in Figure 5.09.

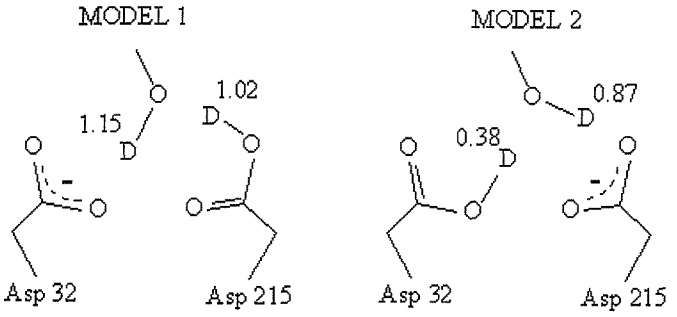


Figure 5.09 Showing the refined occupancy values for the deuteriums in the two different models of the active site.

The differences in the occupancy refinement between the two models show the occupancy for a deuterium to be three times as high for Asp 215 compared to Asp 32. They also confirm that the hydrogen on the hydroxyl group of LOV 405 has exchanged for deuterium. A number of aspartic proteinases including endothiapepsin have low pI values. It has been suggested that these low pI values are due to buried carboxylate groups that remain deprotonated even at very low pH. Evidence for this can be seen in the neutron model in which a number of buried carboxylates appear to be deprotonated. In general the carboxylate groups of these residues interact with buried main chain >N-H groups or polar side chain atoms. While a number of these make salt bridge interactions and therefore will not contribute to the net charge of the molecule, some do not and instead interact with neutral polar groups, which may function to stabilise the negative charge on the carboxylates. A definitive example of a buried negatively charged carboxylate is Asp 87 which is an almost completely conserved residue. The side chain of this residue accepts hydrogen bonds from the side chains of Tyr 56, Ser 61 and Thr 63 as well as the main chain >N-H of Thr 88 (Figure 5.10). The neutron density clearly shows that these four groups are deuterated which strongly indicates that the aspartate is negatively charged. Of the residues making these interactions, the Ser and Thr residues are strongly conserved in the aspartic proteinases; the Tyr residue is also conserved but to a lesser extent.

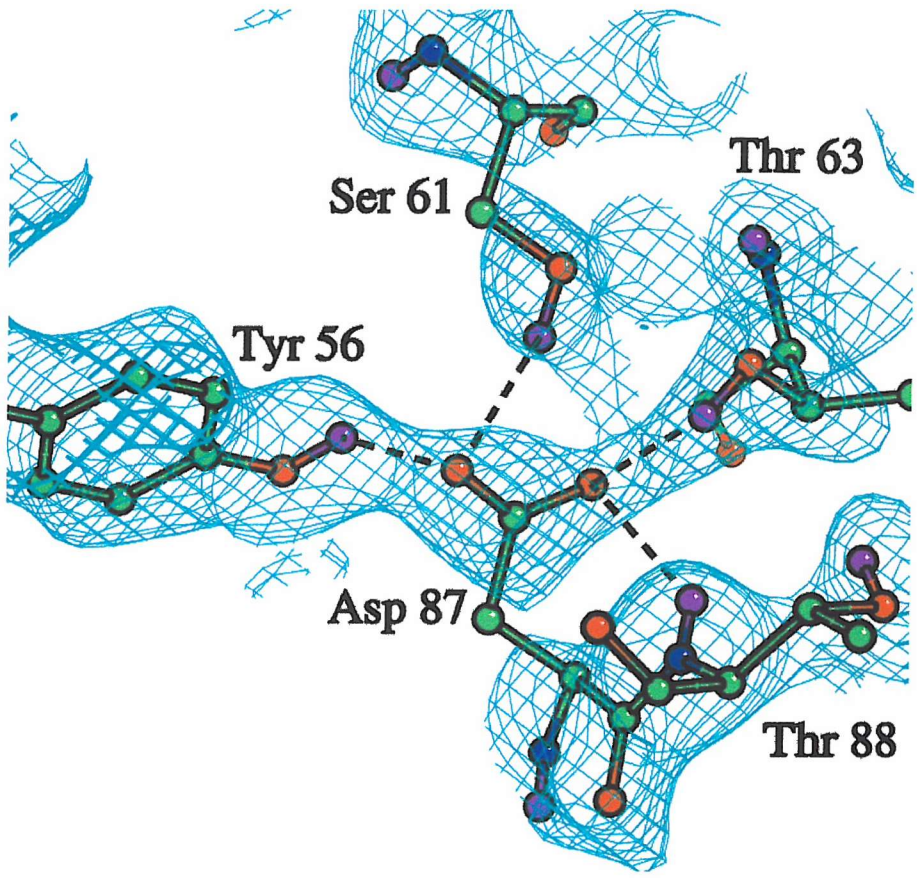


Figure 5.10 Showing the hydrogen bonding pattern of Asp 87 a negatively charged aspartate. The $2mF_o$-DF_c density is shown in cyan at 1.2 σ.

X-ray Data collection of Endothiapepsin inhibitor complexes

All the X-ray data were collected at european synchrotron light radiation sources, the data for the H189 inhibitor complex was collected at HASYLAB while the rest of the data was collected at the ESRF. The X-ray detector used at HASYLAB was a Mar Research CCD while an ADSC Quantum 4 CCD was utilised at the ESRF.

PD-130,328, CP-80,794, PD-129,541, H256 and PD-135,040 complex data collection

Monochromatic X-ray diffraction data were collected at the ESRF Grenoble beamline ID 14-2 on four endothiapepsin inhibitor complexes (PD-130,328, PD-129,541, H256 and CP-80,794). This beamline has a fixed wavelength of 0.933Å and a beam size of approx 50-200 microns and is fitted with an ASDC Q4 CCD detector with a readout time of approximately. five seconds. An Oxford Cryostream was used for the cryo-cooling of all crystals during data collection. Due to the strong diffraction obtained from all of the crystals, overloads were present on all the high resolution (0.86Å) data sets. Thus medium (2.0Å) and low resolution (3.5Å) data sets were collected for all complexes by attenuating the beam and increasing the sample to detector distance to ensure that the low resolution reflections were properly recorded. This gave rise to three data sets for each crystal but ensured that the maximum number of reflections where correctly recorded. The oscillation angle for the high resolution data pass was 0.5° while the medium and low resolution data passes used an oscillation angle of 1° with exposure times of 3, 3 and 6 seconds respectively. For each low resolution pass the incident beam was attenuated by a factor of 2.5 by the inline attenuation system to ensure that all intensities are within the dynamic range of the detector.

The program MOSFLM (Leslie 1992) was used to autoindex the first diffraction pattern from each of the crystals and after the correct unit cell parameters had been selected the unit cell orientation was refined. The strategy option of MOSFLM was then invoked which calculated the ϕ region for data collection that would give the maximum data completeness. These starting and finishing ϕ angles were then entered into the ProDC software, which controls the X-ray machinery in the experimental hutch. The data on the endothiapepsin PD-135,040 complex was collected in the same manner at the ESRF on beamline ID 29 using the installed ADSC Q210 CCD detector and a wavelength of 0.91 Å. A high resolution pass of 180° of data was collected with an oscillation angle of 1° at a distance of 180 mm with an exposure time of 1 second. A low resolution pass of 200° of data was also

collected with an oscillation angle of 2° at a crystal to detector distance of 200 mm. The beam was attenuated by a factor of 2.5 using the in line attenuation system. The exposure time for this pass was 6 seconds due to the attenuation of the beam.

The procedure used to collect the data on the endothiapepsin H189 crystal at HASYLAB was the same as outlined above apart from the fact that it was only necessary to collect a high (0.96Å) and low resolution data set (3.0 Å) both with an oscillation angle of 1° with exposure doses of 2000 and 100 kilo counts respectively.

Endothiapepsin PD-130,328 complex refinement

After the generation of the SHELX ins file containing the atomic positions from the 3ER5 pdb file (endothiapepsin H189 complex), inhibitor restraints and refinement parameters a structural refinement with SHELX was started. In the initial refinement the protein was treated as a rigid body and refined against the data to a resolution limit of 2.5 Å to account for any differences in the gross position of the protein in the unit cell. Refinement then shifted to an isotropic conjugate gradient least squares (CGLS) algorithm which produced a structure with an R_{factor} of 21.0 % and an R_{free} of 22.0 %. After this the fit of the protein in the electron density was checked for the entire molecule. Alterations to the conformation of a number of side chains were made. The conformations of a large number of serine and lysine side chains were altered to better fit the electron density.

The refinement of the water structure then took place using SHELXWAT, a program for automatic divining of water molecules in macromolecular structures. The default settings were used enabling 10 cycles of refinement with a maximum of 50 water molecules to be added per refinement cycle. This refinement took 43 hours of computer time adding a total of 212 water molecules to the structure, taking the total number of water molecules to 451. At the end of the refinement

this gave an R_{factor} of 18.0 % and an R_{free} of 20.0 %. The water structure was then checked over by eye using XTALVIEW (McRee 1999) and after some waters had been moved or removed, the fit of the protein to electron density was then checked over again. Where it became clear that a residue present on the surface of the protein had a dual occupancy, a pdb file was created defining the fit of a second conformer in the extra electron density and this was entered into SHELX for occupancy refinement. During refinement a number of surface side chains were identified as being in two conformations. These where Ser 42, Glu 44, Asp 114, Ser 145, Val 150, Ser 204, Ser 230, Val 252 and Ser 263. The total number of disordered residues is therefore nine, representing only 2.73 % of all residues in the structure which is the same figure calculated for the H189 structure although there are some differences in which side chains are disordered. The ins file was then edited to make use of anisotropic B factor refinement via the SHELX command flag ANIS. This enables six parameters to be used to define atomic movement, increasing the accuracy of the model. This refinement resulted in a model with an R_{factor} 14.0 % of and an R_{free} of 16.0 %. The structure was then checked over and the water structure remodelled to take into account changes in the Fourier maps and the density around the multiple conformations was also checked. In the next stage of refinement, riding hydrogens were added to the model after which the refinement converged producing a structure with the crystallographic statistics shown in Table 4 (values for the outer resolution shell are given in brackets). The Ramachandran plot for the final structure in which no outliers were detected is shown in Figure 5.11.

PDB code	1gvw
Unit Cell	A 43.88 Å, b 75.45 Å, c 43.23 Å, β 97.4°
Space Group	$P2_1$
Number of unique reflections	141,699
Resolution Range Å	10-1.00
Outer Shell Å	1.05-1.00
Multiplicity	5.0 (3.5)
Mean I/σ(I)	3.0 (2.0)
R_{merge}	10.0% (26.3%)
Data completeness	100% (100%)
R_{factor}	12.07%
R_{free}	14.40%
Number of water molecules	457
Average B_{iso} protein atoms	13.9
Average anisotropy protein atoms	0.66
Data to parameter ratio	6.30

Table 4 Crystallographic statistics for the endothiapepsin PD-130,328 structure. Figures for the outer shell are given in brackets.

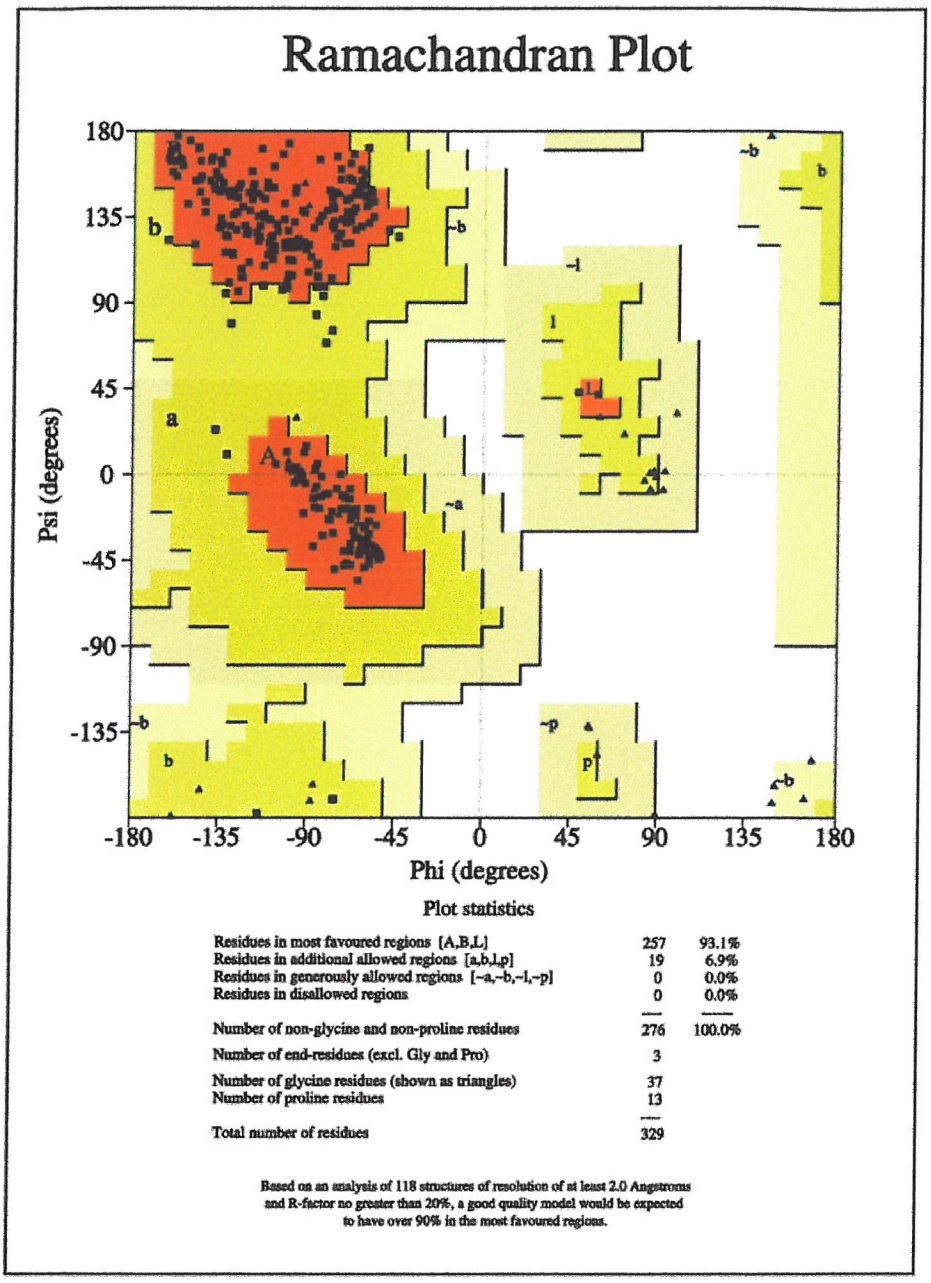


Figure 5.11 A Ramachandran plot generated from the final refinement cycle of the endothiapepsin PD-130,328 complex.

After the refinement had converged the electron density around the active site was checked and the $2mF_o$-DF_c electron density contoured at 1.0 σ around the active site is shown in Figure 5.12.

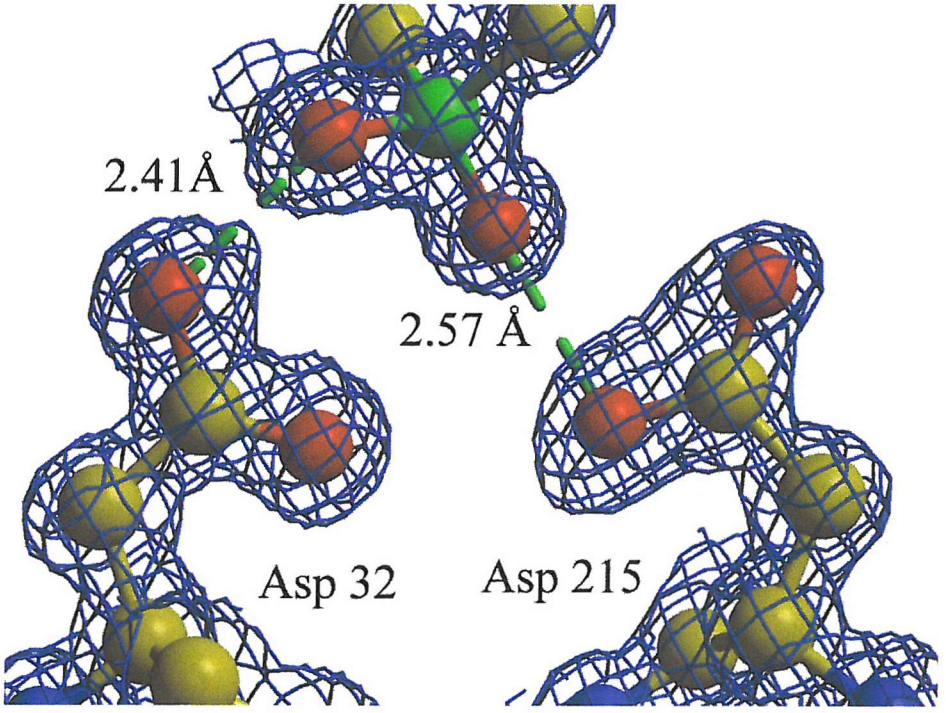


Figure 5.12 The electron density in the endothiapepsin PD-130,328 complex active site. The $2mF_o$-DF_c density is shown in blue.

Then a refinement of the model took place with all stereochemical restraints dropped for the aspartate and glutamate carboxyls. This yielded a structure with an R_{factor} of 12.50 % and an R_{free} of 15.0 %. Following this, the model was refined against all data (including the R_{free} set) and then a full matrix inversion calculation that due to computer limitations was calculated without the anisotropic displacement parameters was performed. All restraints and shift dampeners were removed before the start of the calculation. Then the bond lengths of all the aspartate and glutamate residues in a single conformation were measured to verify that the protonation states of these residues could be determined. As endothiapepsin has optimal activity at low pH it has been speculated that some of the ionisable side chains remain unprotonated at low pH. This theory was tested by checking the bond lengths of the ionisable groups and calculating the difference between them. This was compared with the standard deviation for the bond length difference. If the bond length difference was greater than two and a half times the

ESD variance the aspartate or glutamate was classified as protonated. If the bond length difference was less then two and a half times the ESD variance then the aspartate or glutamate was classified as charged. This analysis was carried out on all of the atomic resolution X-ray structures. As can be seen from the PD-130,328 phosphinic acid model shown in Figure 5.13 the overall protonation states of the catalytic aspartates remain the same as in the Veerapandian *et al* 1992 mechanism. However the inner oxygen OD1 of Asp 215 would appear to be protonated when the PD-130,328 acid group binds. In the Veerapandian mechanism it is the OD2 of Asp 215 that is protonated in the transition state complex. This change in protonation state of Asp 215 probably occurs because of the binding of the bulky phosphinate in the active site, which is reflected in the high K_i value for this inhibitor (110nm). The bulky phosphinate group probably also causes the very short hydrogen bonds present in the structure between the outer oxygen of Asp 215 and the phosphinic O1 (2.57 Å) atom and between the inner oxygen of Asp 32 and the phosphinic O2 (2.41 Å).

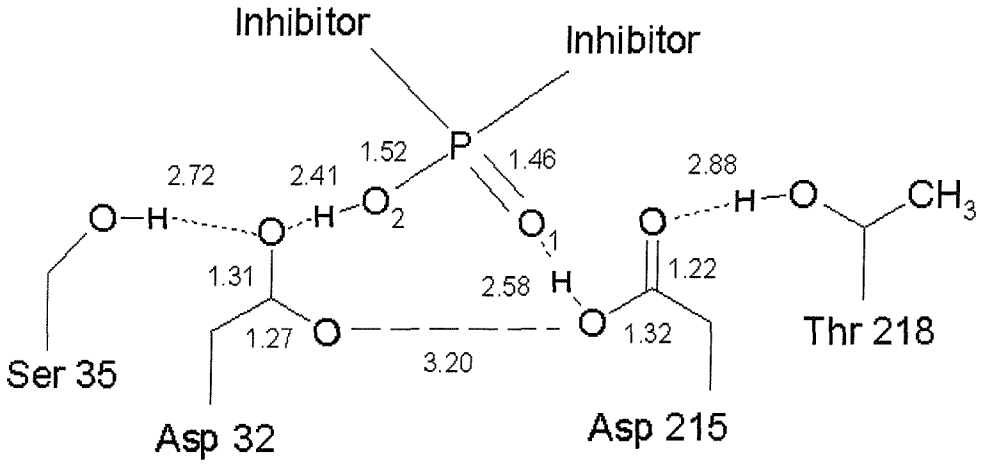


Figure 5.13 PD-130,328 bound to the active site of endothiapepsin, all bond lengths are shown in Ångstroms. The ESD values for the carboxyl bonds on Asp 32 are both 0.0105 Å while for the ESD for Asp 215 are 0.0133 Å to the inner oxygen and 0.014 Å to the outer oxygen.

The 50% thermal elliposoids in the active site of the PD-130,328 complex are shown in Figure 5.14. They are fairly large for atoms in an atomic resolution structure.

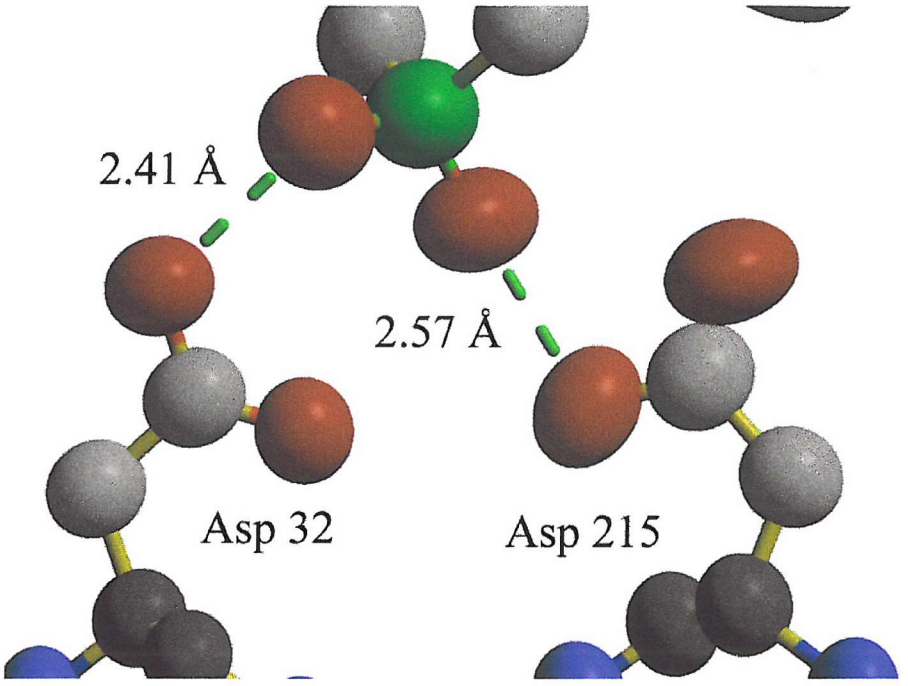


Figure 5.14 Showing the 50% probability thermal ellipsoids for the atoms at the PD-130,328 active site. The ADPs are fairly large and isotropic for an atomic resolution structure.

In the PD-130,328 complex it should be noted that Asp 215 is displaced in the active site when the structure is compared to the other X-ray structures, it is forced further apart from Asp 32 and pushed outwards as shown in Figure 5.15. The same phenomenon was observed when the structure was solved at lower resolution (Lunney *et al* 1993).

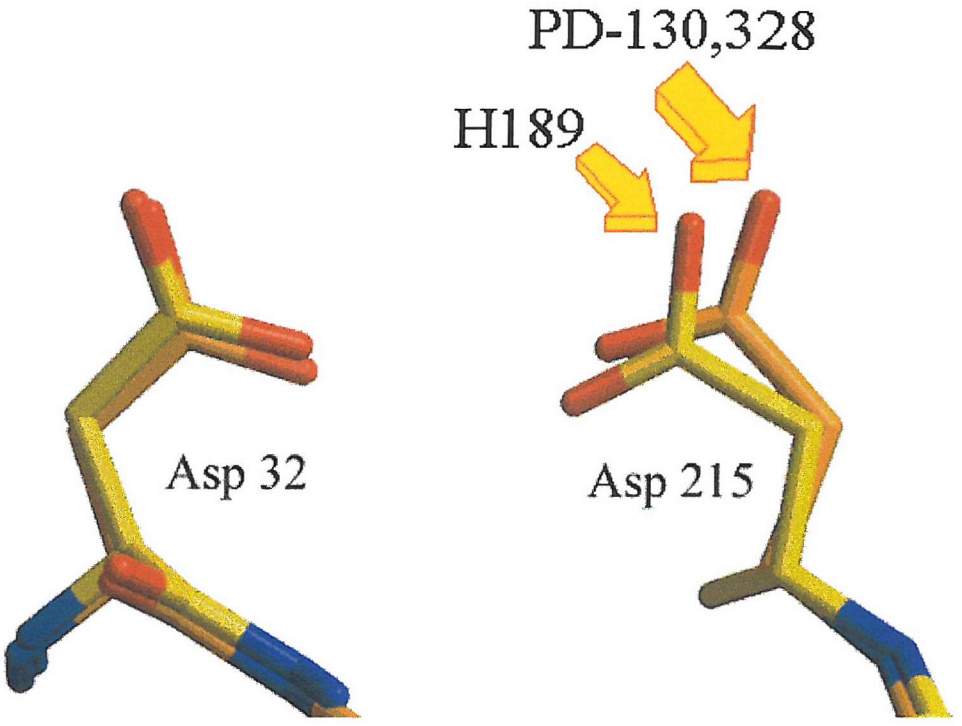


Figure 5.15 Showing the displacement of Asp 215 in the PD-130,328 structure compared to the H189 structure. The H189 structure is shown in yellow with the PD-130,328 structure shown in orange.

Endothiapepsin H189 Complex refinement

The H189 complex was refined using the same methods as the PD-130,328 acid complex. Before any refinement started 5 % of the reflections where flagged using SHELXPRO to form the R_{free} set. After the R_{free} reflection set had been generated the general fit of the protein to the $2mF_o$-DF_c density was checked and the water structure refined and two sulphate ions modelled into the structure. Following this multiple conformations for some side chains were fitted where the mF_o-DF_c map indicated positive density suggesting two or more conformations for a side chain. After a number of rounds of refinement to ensure all water positions or multiple conformations had been found, the model was then refined anisotropically which caused the R_{factor} to drop from 15.10 % to 11.90 % and the R_{free} to drop from 16.90

% to 13.40 %. The protein and water structures were then examined again and remodelled before riding hydrogens where attached to the C,N,O atoms. A number of side chains in the structure were modelled in different conformations these included Ser 42, Glu 44, Asp 114, Ser 151, Val 150, Ser 204, Ser 206, Ser 263 and Ser 279. This was done and the occupancies refined for each conformation, all of the residues with multiple conformations were located on the surface of the molecule. The total number of disordered residues is therefore nine representing only 2.73 % of all residues in the structure. The refinement converged producing a structure with the crystallographic statistics shown in Table 5 (values for the outer resolution shell are given in brackets).

PDB code	1gvu
Unit Cell	a 42.48 Å, b 75.78 Å, c 42.99 Å, β 95.43 °
Space Group	$P2_1$
Number of unique reflections	146,980
Resolution Range Å	10-0.935
Outer Shell Å	0.99-0.94
Multiplicity	2.7 (2.1)
I/σ(I)	11.50 (2.8)
R_{merge}	4.00% (16.3%)
Data completeness	86.2% (72.5%)
R_{factor}	11.06%
R_{free}	13.29%
Number of water molecules	462
Average B_{iso} protein atoms	8.47
Average anisotropy protein atoms	0.55
Data to parameter ratio	5.31

Table 5 Crystallographic statistics for the endothiapepsin H189 complex. Figures for the outer shell are given in brackets.

All residues in the structure are in the favoured or additionally allowed regions of the Ramachandran plot (Figure 5.16).

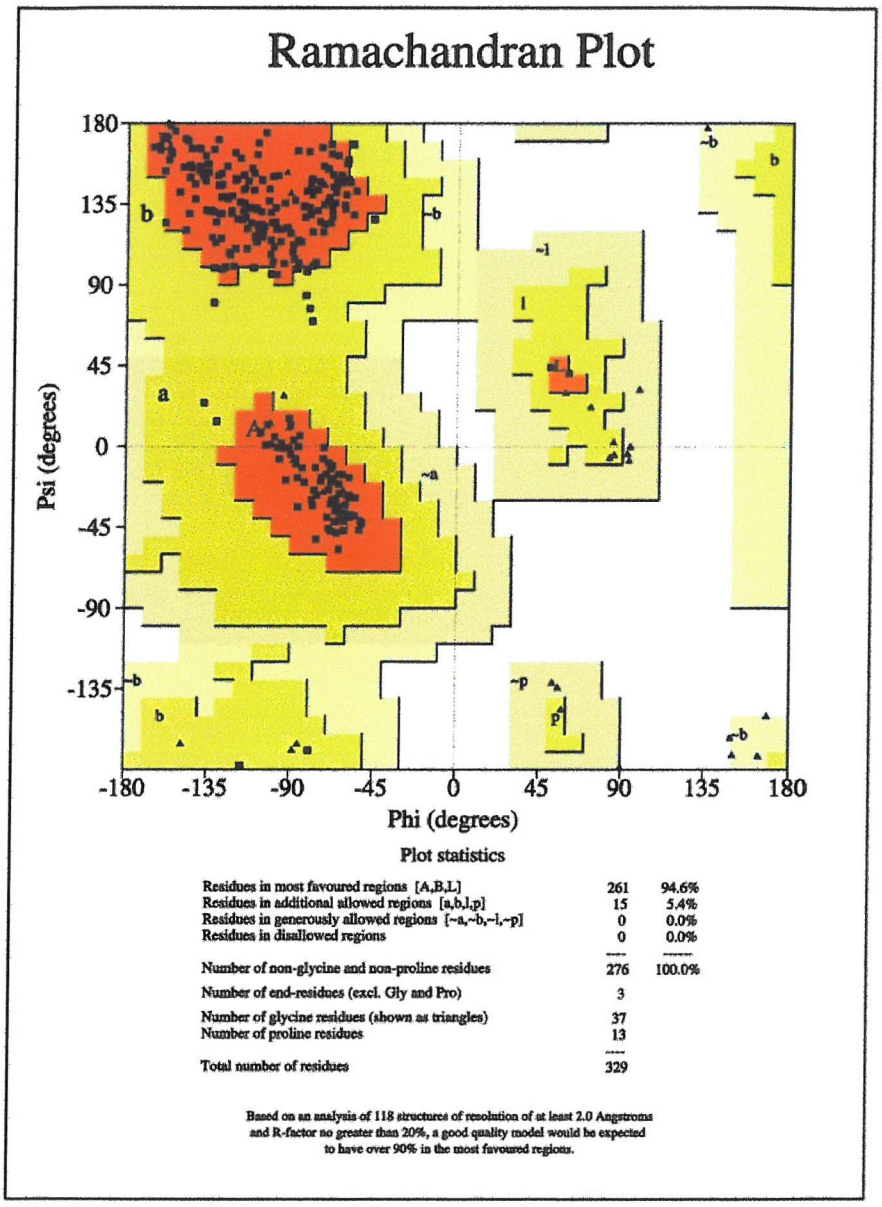


Figure 5.16 A Ramachandran plot generated from the final refinement of the endothiapepsin H189 structure.

After the refinement had converged unrestrained refinement of the aspartate and glutamate carboxyls was performed to determine the C-O bond lengths. The model was then refined against all data (including the R_{free} set) before a full matrix inversion without ADPs was calculated for the entire protein with all stereochemical restraints and shift dampeners removed. The bond lengths and ESDs for the catalytic aspartates were then checked and are shown in Figure 5.17.

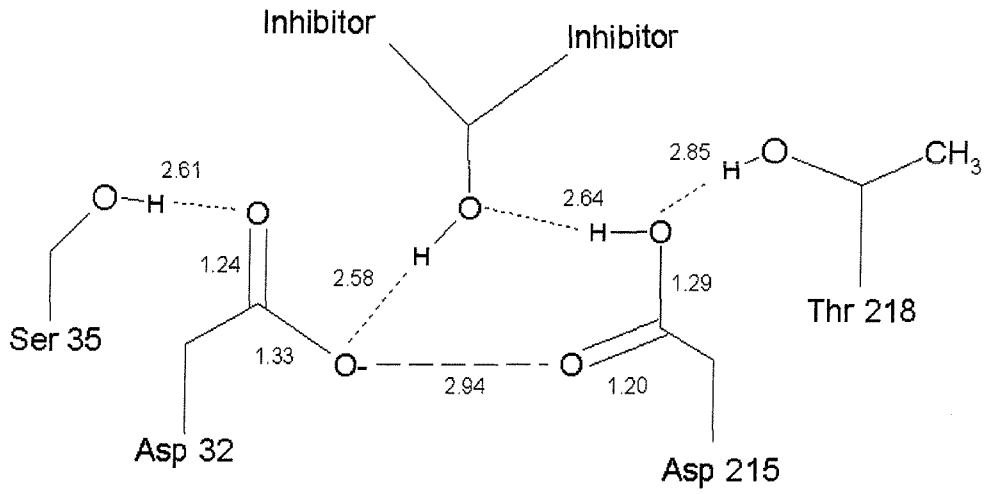
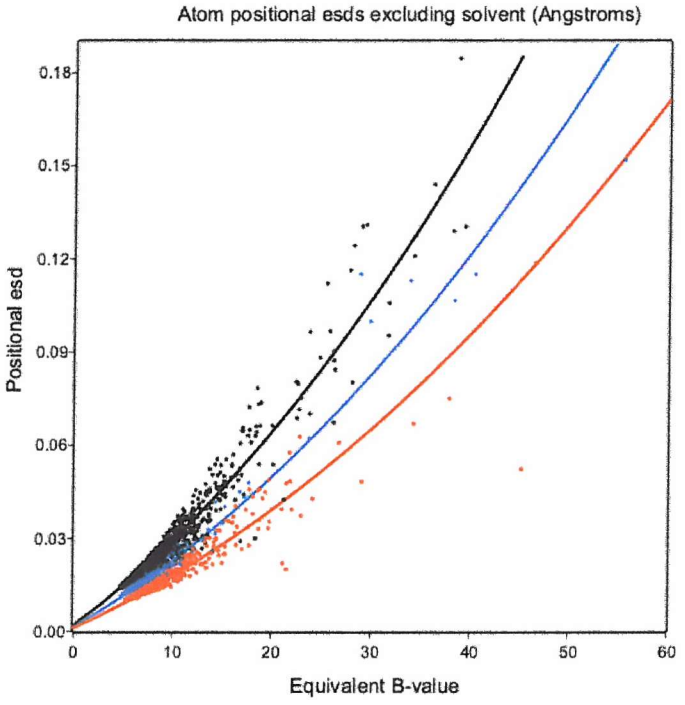


Figure 5.17 H189 bound to the active site of endothiapepsin. All bond lengths are shown in Å. The ESDs for all four aspartate C-O bonds are 0.01 Å.

The values for all atom positional ESDs for all protein atoms are shown in Figure 5.18a. The ESD values of the C-C, C-O, C-N bond lengths are given in figure 5.18b.

(a)



(b)

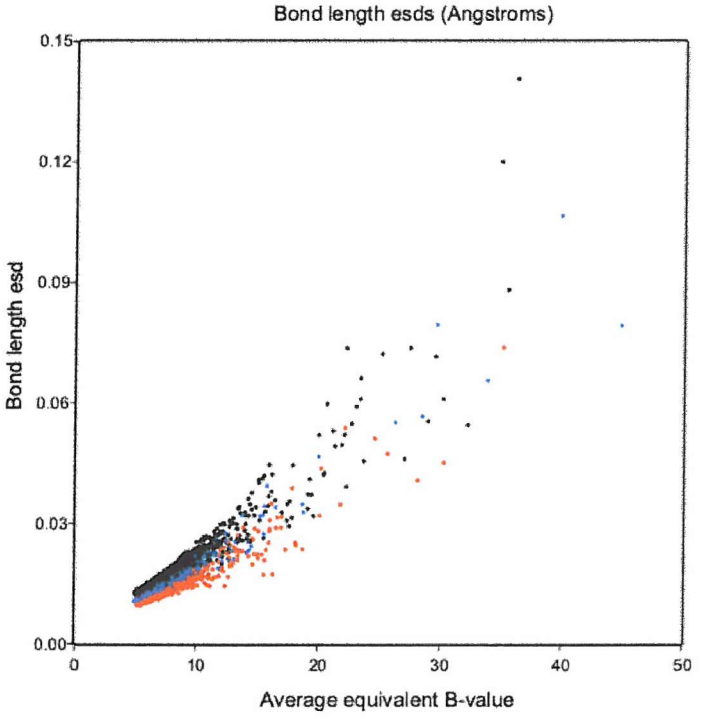


Figure 5.18 The results of an unrestrained least squares matrix for the endothiapepsin H189 structure, all protein atoms and bond lengths are shown there are no outliers. In (a) carbon atoms are represented in black nitrogen in blues and oxygen in red while in (b) C-C bonds are represented in black C-N bonds in blue and C–O, C=O bonds in red.

The greater accuracy in the positions of the oxygen atoms compared with nitrogen and nitrogen compared with carbon within the structure can clearly be seen for atoms with B_{iso} values < 20; this trend was also observed in Deacon *et al* (1997). This trend is also valid for bond length accuracies thus increasing the validity of the bond lengths and ESD values obtained for the catalytic aspartates. In can be seen in Figure 5.18 that many bond lengths with an average B_{iso} < 10 have ESD values of less than 0.014 Å. For these bonds the diffraction data has a greater weight than the stereochemical dictionaries (Cruickshank 1999). The catalytic aspartates and Ser 35 all have B_{iso} values less than 10 and bond length ESD values less than 0.014 Å indicating that their bond lengths are mainly derived from the diffraction data.

Endothiapepsin CP-80,794 Complex refinement

This inhibitor complex was refined in the same manner as the other high resolution X-ray structures. Once again a number of side chains in the structure could be modelled in different conformations these were Ser 42, Glu 44, Ser 90, Asp114, Leu 120, Ser 126, Ser 131, Ser 145, Val 150, Thr 176, Ser 206, Ser 208, Thr 227, Ser 231, Ser 263, Ser 279 and Leu 321. This was done and the occupancies refined for each conformation. The total number of disordered residues is 17 representing 5.15 % of all residues in the structure. The refinement converged with the crystallographic statistics shown in Table 6 (values for the outer resolution shell are given in brackets).

PDB code	1gvt
Unit Cell	a 42.55 Å, b 74.62 Å, c 44.43 Å, β 97.0 °
Space Group	$P2_1$
Number of unique reflections	149,856
Resolution Range Å	10-0.98
Outer Shell Å	1.03-0.98
Multiplicity	4.4 (3.4)
$I/\sigma(I)$	5.20 (2.3)
R_{merge}	7.4 % (15.8 %)
Data completeness	99.5 % (100 %)
R_{factor}	11.02 %
R_{free}	13.14 %
Number of water molecules	473
Average B_{iso} protein atoms	6.94
Average anisotropy protein atoms	0.48
Data to parameter ratio	5.11

Table 6 Crystallographic statistics for the endothiapepsin CP-80,794 structure. Figures for the outer shell are given in brackets.

There were no outliers observed in the Ramanchdran plot (Figure 5.19).

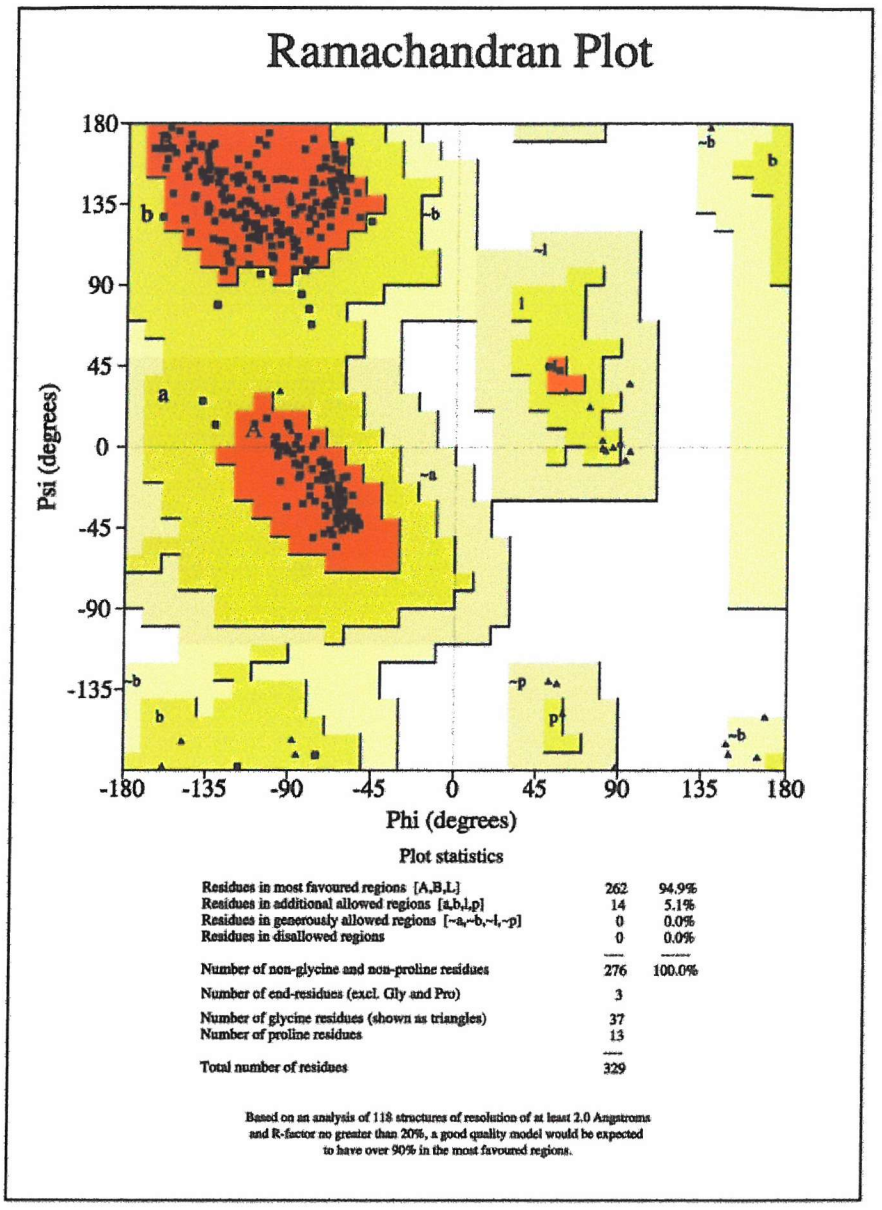


Figure 5.19 A Ramachandran plot generated after the final refinement of the endothiapepsin CP-80,794 structure.

After the refinement had converged the restraints on the C-O bond lengths of the aspartates and glutamates were then dropped and these bond lengths where then refined. The structure was then refined against all data including the R_{free} set with all stereochemical restraints and shift dampeners removed and the ESDs for all the bond lengths and atomic positions in the structure where calculated by a full

matrix inversion with the ADPs omitted. The bond lengths and ESDs for the catalytic aspartates were then checked and are shown in Figure 5.20. The ESDs for the Asp 32 bonds are 0.01 Å and 0.011 Å for Asp 215.

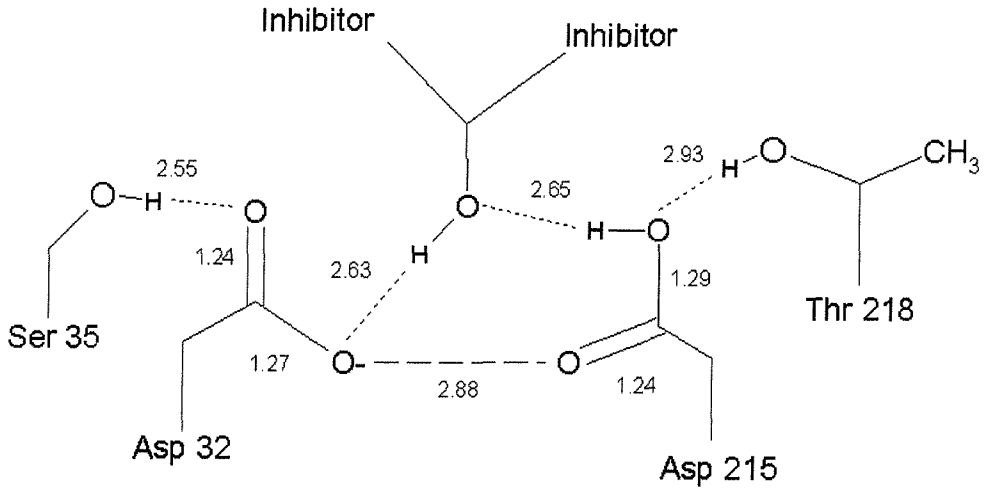


Figure 5.20 Showing the bond lengths in the active site of endothiapepsin bound to CP-80,794 in Å. The bond length ESDs for Asp 32 are both 0.010 Å and 0.011 Å for Asp 215.

Inspection of the active site region revealed excellent electron density for one of the hydrogens in the active site, the electron density maps are shown in Figure 5.21 in which the $2mF_o$-DF_c density at 1 σ is coloured blue, the $2mF_o$-DF_c density is coloured green at +2.5 σ and yellow at -2.5 σ (Coates *et al* 2002).

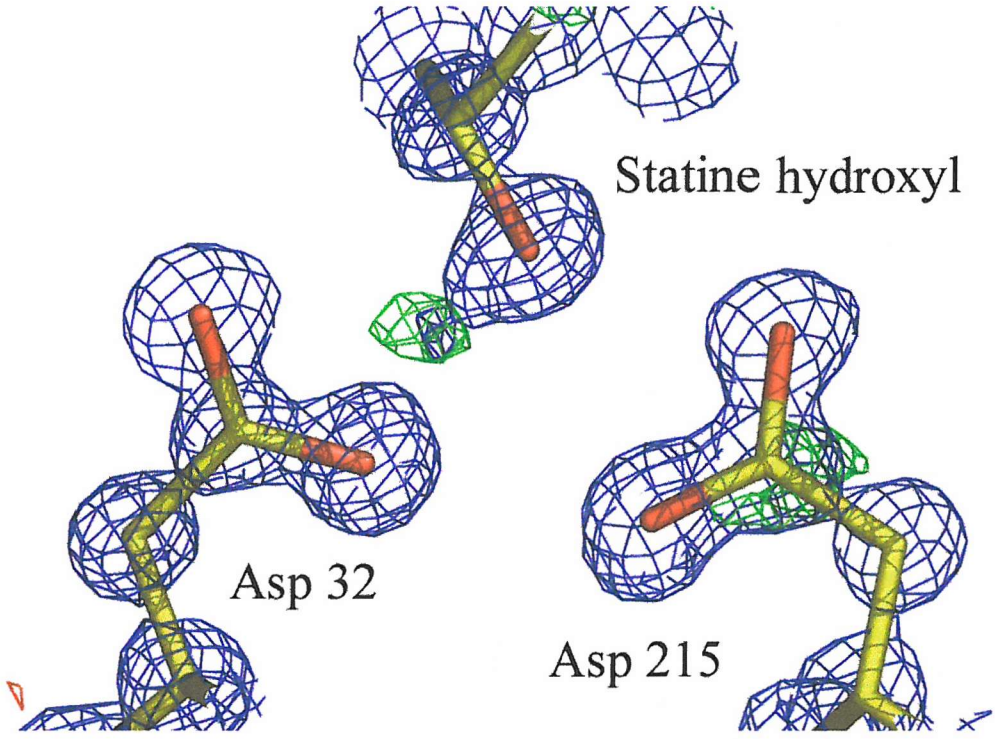


Figure 5.21 Showing the electron density around the active site of CP-80,794. The $2mF_o$-DF_c density at 1 σ is coloured blue, the mF_o-DF_c density is coloured green at +2.5 σ and red at -2.5 σ. There is excellent density for the hydrogen on the statine hydroxyl oriented towards the inner oxygen of Asp 32.

There is clear positive $2mF_o$–DF_c and mF_o-DF_c density for a hydrogen atom on the statine inhibitor hydroxyl orientated towards the inner oxygen of Asp 32 suggesting the presence of a short hydrogen bond between the two oxygen atoms. There is also electron density at + 2 σ in the mF_o-DF_c electron density map for the presence of a hydrogen atom on the outer oxygen of Asp 215 however at this contour level the maps become somewhat noisier. The 50 % thermal ellipsoids for the atoms at the active site are shown in Figure 5.22; they are fairly small and isotropic indicating that the active site region is well ordered. The ellipsoid for the oxygen atom in the inhibitory statine group points towards the inner oxygen of Asp 32 suggesting the formation of a hydrogen bond between these two atoms. As the pK_a of a hydroxyl group is six pH units higher than that of carboxyl group, at the crystallisation pH of 4.5 the hydroxyl group is likely to be protonated

indicating the inner oxygen of Asp 32 in unprotonated when the inhibitor is bound. When a hydrogen atom is modelled into the electron density the C-O-H bond angle is 109.1°, which is very close to the ideal bond angle of 110°. The O-H bond length for this hydroxyl is 1.24 Å, which is longer than the expected value of 1.00 Å. However this O-H bond elongation is a known effect of LBHBs (Clealand *et al* 1998). The separation between the inner oxygen of aspartate 32 and this hydrogen atom is 1.44 Å.

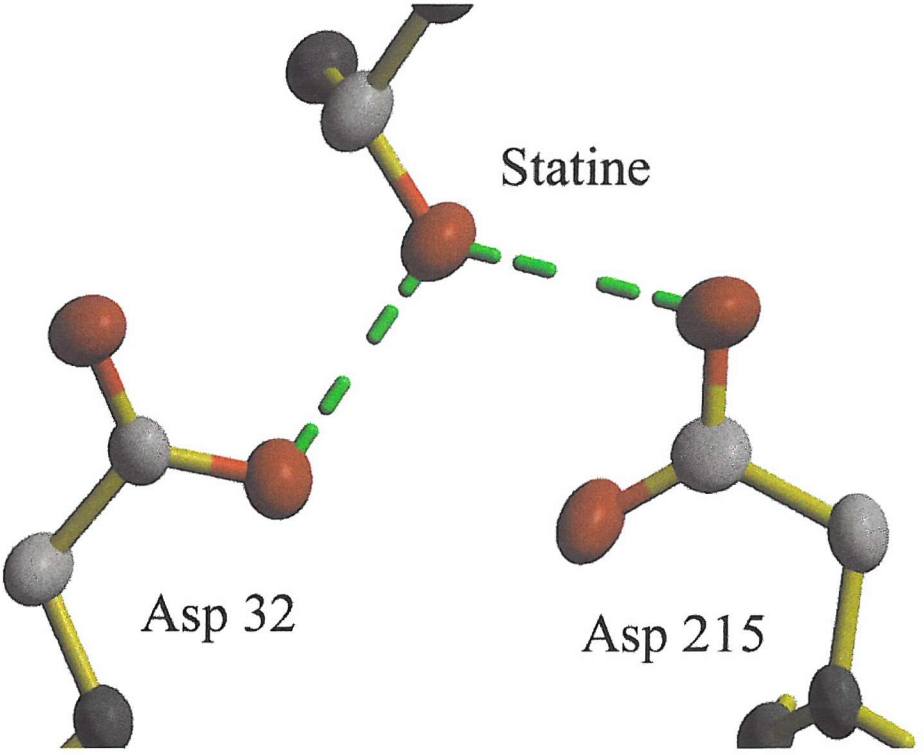
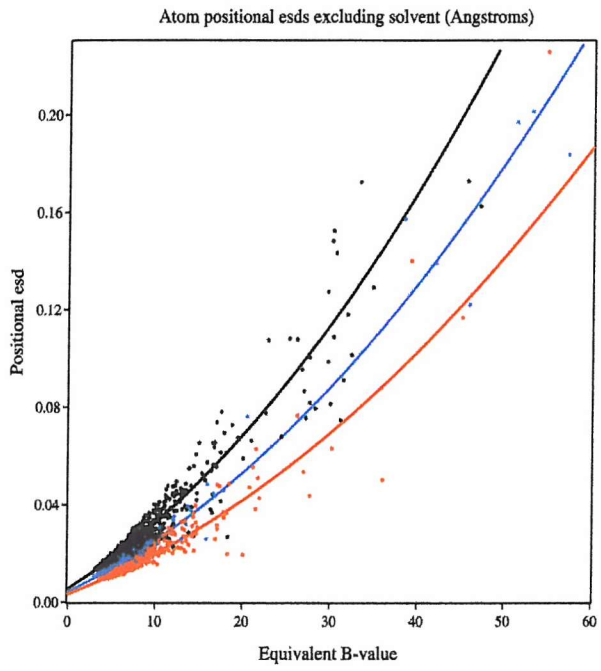


Figure 5.22 Showing the 50 % probability thermal ellipsoids for the active site of CP-80,794. The axis of the ellipsoid representing the statine hydroxyl is orientated towards the inner oxygen of Asp 32.

The mean anisotropy for all protein atoms is 0.46 with a standard deviation of 0.15, which is very close to the mean anisotropy values calculated in Merritt 1999a of 0.45 with a standard deviation of 0.15. These values were obtained using the standard SHELX defaults for SIMU and ISOR, and are slightly different to the anisotropy values calculated for the H189 structure. An unrestrained full matrix inversion was performed on the CP-80,794 structure using SHELX 97-2 (Sheldrick 1998); due to computer limitations the anisotropic displacement

parameters were omitted from the calculation. The positional atomic ESD and bond length ESD plotted against B_{iso} for the unrestrained full matrix calculation are shown in Figure 5.23.

(a)



(b)

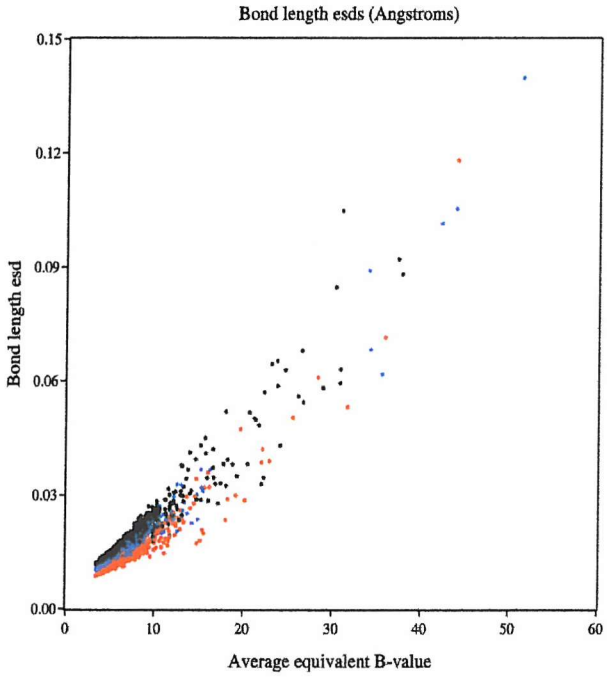


Figure 5.23 The results of an unrestrained least squares matrix for the endothiapepsin CP-80,794 structure, all protein atoms and bond lengths are shown and there are no outliers. In (a) carbon atoms are represented in black nitrogen in blues and oxygen in red while in (b) C-C bonds are represented in black C-N bonds in blue and C–O, C=O bonds in red.

The greater precision in the position of oxygen atoms is clear from the Figure 5.23a and translates to the C-O and C=O bonds having the lowest average ESD. These results are in line with the values obtained from proteins refined at similar resolution (Deacon *et al* 1997). The lowest bond length ESD in the entire structure is for the β carbon to γ oxygen sidechain bond in Ser 35 (0.0088 Å). This coupled with the short hydrogen bond length to the outer oxygen of Asp 32 suggests that this residue helps stabilise the conformation of Asp 32 when the tetrahedral intermediate is bound to the enzyme.

Endothiapepsin PD-129,541 Complex refinement

This inhibitor complex was refined in the same manner as the other high resolution X-ray structures. During refinement the side chains of the following residues where modelled in two alternative conformations and the occupancy of each conformation refined Thr 7, Ser 42, Glu 44, Ser 74, Asp 114, Ser 126, Ser 145, Val 150, Ser 263 and Ser 279. These ten residues are all found on the surface of the protein and represent 3.03 % of the amino acids in the protein. The refinement converged with the statistics shown in Table 7 (values for the outer resolution shell are given in brackets).

PDB code	1gvv
Unit Cell	a 42.47 Å, b 74.31 Å, c 42.81 Å, β 97.6 °
Space Group	$P2_1$
Number of unique reflections	162,992
Resolution Range Å	10-1.05
Outer Shell Å	1.11-1.05
Multiplicity	4.9 (3.7)
$I/\sigma(I)$	5.30 (1.7)
R_{merge}	7.10% (14.0 %)
Data completeness	98.3% (99.4 %)
R_{factor}	11.63 %
R_{free}	14.04 %
Number of water molecules	492
Average B_{iso} protein atoms	7.79
Average anisotropy protein atoms	0.50
Data to parameter ratio	5.09

Table 7 Crystallographic statistics for the endothiapepsin PD-129,541 complex. Figures for the outer shell are given in brackets.

No outliers were detected in the Ramachandran plot of the final structure (Figure 5.24).

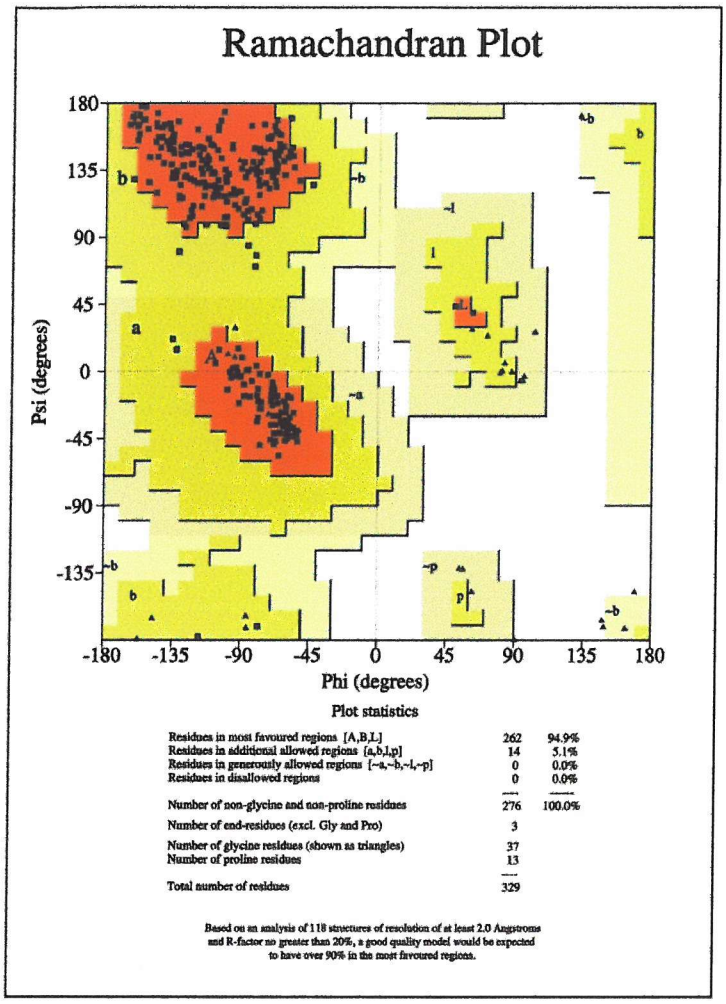


Figure 5.24 A Ramachandran plot produced after the final refinement of the endothiapepsin PD-129,541 complex.

After the refinement had converged the restraints on the C-O bond lengths of the aspartates were then dropped and these bond lengths where then refined. The structure was then refined against all data including the R_{free} set and the ESDs for the C-O bond lengths of the catalytic aspartates were then calculated from a full matrix inversion without ADPs. The bond lengths and ESDs for the catalytic aspartates were then checked and are shown in Figure 5.25.

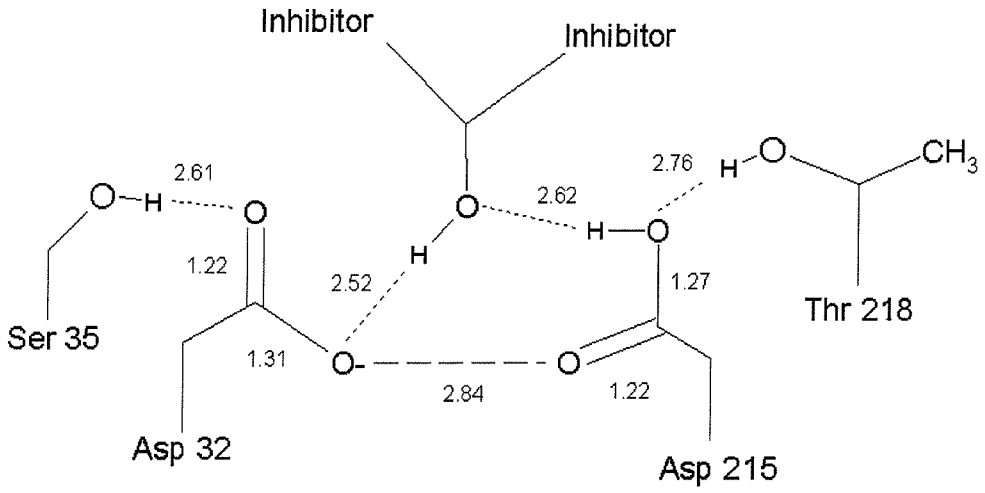
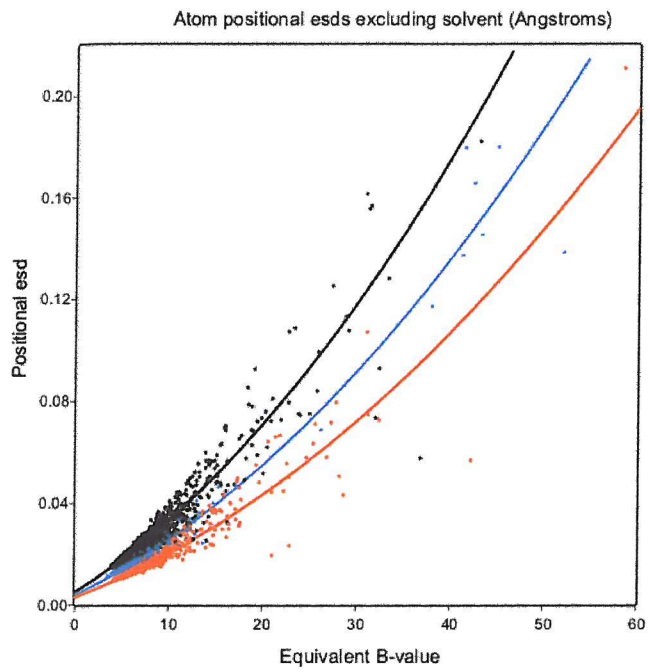


Figure 5.25 showing the PD-129,541 inhibitor bound to the active site, all bond lengths are shown in Å. The bond length ESDs for the Asp 32 bonds are both 0.010 Å and 0.011 Å for Asp 215.

During refinement of the protein structure discrete positive and negative 2σ mF_o-DF_c density was observed around the disulphide bridge between residues cysteine 250 and cysteine 283. To model the mF_o-DF_c density both of the cysteine residues were modelled as discreetly disordered between two sites and their occupancies refined and the restraints relating to disulphide bridge removed. After refinement the position and occupancy of the conformers were checked. The results of the occupancy refinement confirmed the fact that the disulphide bond has indeed broken. However no convincing alternative conformations could be found for the sulphur atoms in the disulphide bridge. The occupancy for the sulphur atom of cysteine 250 was 0.63 and 0.81 for cysteine 283. The atom positional and bond length ESDs for all protein atoms are shown in Figure 5.26.

(a)



(b)

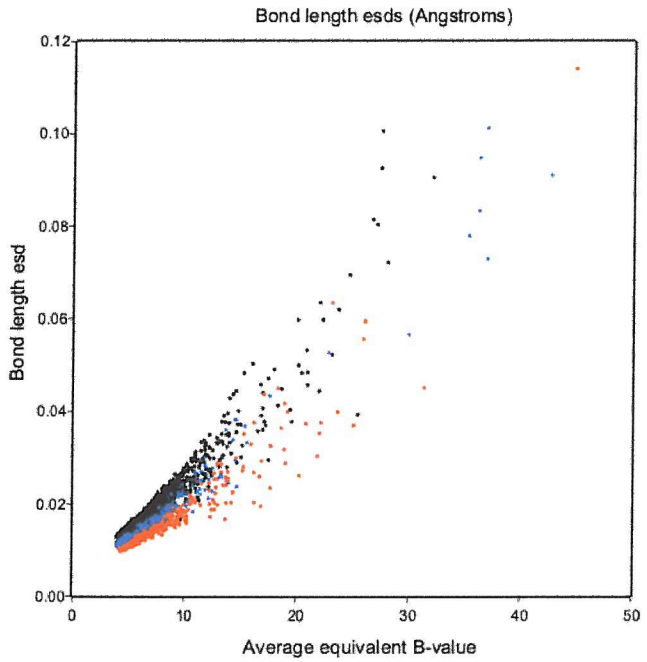


Figure 5.26 The results of an unrestrained least squares matrix inversion for the endothiapepsin PD-129,541 structure, all protein atoms and bond lengths are shown; there are no outliers. In (a) carbon atoms are represented in black nitrogen in blues and oxygen in red while in (b) C-C bonds are represented in black C-N bonds in blue and C–O, C=O bonds in red.

Endothiapepsin H256 Complex refinement

This inhibitor complex was refined in the same manner as the other high resolution X-ray structures. During refinement the side chains of the following residues where modelled in two alternative conformations and the occupancy of each conformation refined: Ser 42, Glu 44, Ser 66, Ser 78, Asp 114, Ser 126, Ser 131, Asp138, Ser 145, Val 150, Thr 176, Ser 206, Ser 208, Ser 239, Ser 266, Asp 274, Ser 282 and Ile 301. These eighteen residues are all found on the surface of the protein and represent 5.46 % of the amino acids in the protein. The refinement converged with the statistics shown in Table 8 (values for the outer shell are given in brackets).

PDB code	1gvx
Unit Cell	a 43.88 Å, b 75.44 Å, c 43.23 Å, β 97.47°
Space Group	$P2_1$
Number of unique reflections	136,765
Resolution Range Å	10-1.0
Outer Shell Å	1.05-1.00
Multiplicity	6.9 (6.1)
$I/\sigma(I)$	3.70 (3.6)
R_{merge}	7.90% (16.3%)
Data completeness	96.9% (94.3%)
R_{factor}	13.93%
R_{free}	16.47%
Number of water molecules	410
Average B_{iso} protein atoms	23.70
Average anisotropy protein atoms	0.73
Data to parameter ratio	5.26

Table 8 Crystallographic statistics for the endothiapepsin H256 structure. Figures for the outer shell are given in brackets.

Refinement of this complex did not proceed as smoothly as the other complexes and refinement converged with a higher R_{factor} that of the other complexes. This could be attributed to high average protein B_{iso} values for this structure (23.70) in comparison to the much lower B_{iso} values for the statine inhibitor based structures the highest of which is 8.47. The B_{iso} values for the atoms in the active site of this structure are around 20 meaning that the bond length ESD values are likely to be underestimated as the ADPs were omitted from the full least squares matrix inversion. However after refinement convergence the restraints on the aspartate and glutamate carboxyl bond lengths were removed and refined in an unrestrained manner. The model was then refined against all data including the R_{free} set before the least squares matrix was inverted with all restraints and shift dampeners removed. The bonds lengths and ESDs for the catalytic aspartates were checked and are shown in Figure 5.27.

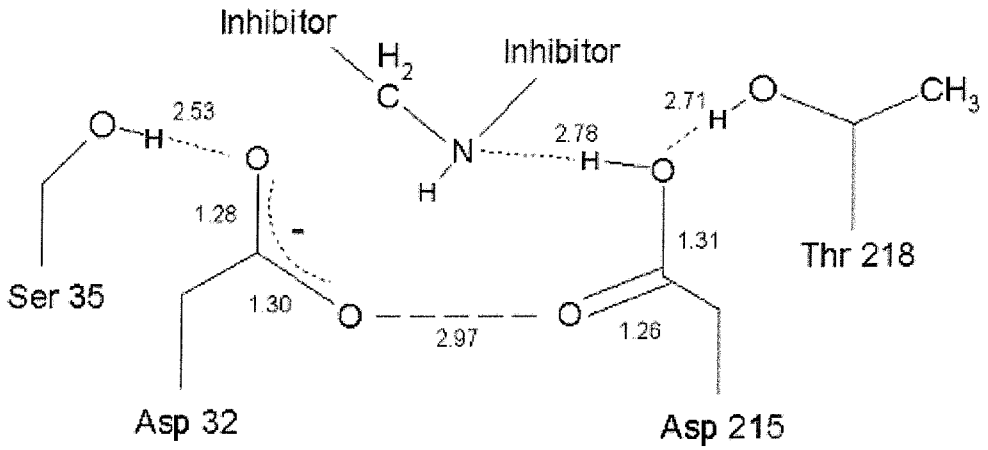


Figure 5.27 Showing the bond lengths in the active site of the H256 structure. The ESDs for all four carboxyl bond lengths is 0.010 Å, however they are likely to be underestimated due to the relatively high B_{iso} values in this structure as the ADPs were omitted from the least squares inversion.

The Ramachandran plot of the final structure showed no unusual Psi-Phi values for any of the amino acid residues in the structure (Figure 5.28).

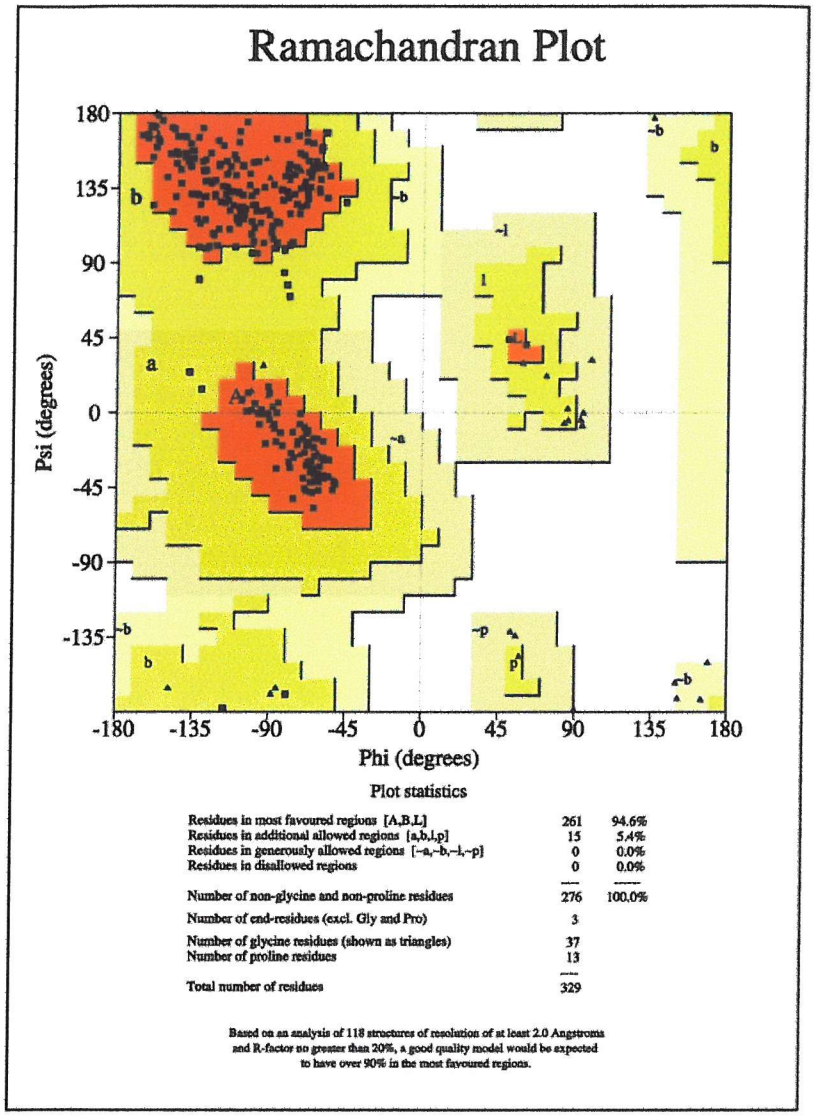


Figure 5.28 A Ramachandran plot of the endothiapepsin H256 structure after the final refinement.

After the refinement of the model was complete the restraint relating to the difference in ADPs between nearby atoms (SIMU) was altered from 0.1 (the default value) to 0.025. The primary effect of this restraint is to make the axes of deformation of nearby atoms roughly parallel (in direction rather than magnitude

of anisotropy). The ISOR restraint which restraints the ADPs of each water atom to be approximately isotropic was also removed. The tightening up of the SIMU restraint has been shown to alter the atomic anisotropy distribution making it more symmetric as shown in Figure 5.29 (Merritt 1999a).

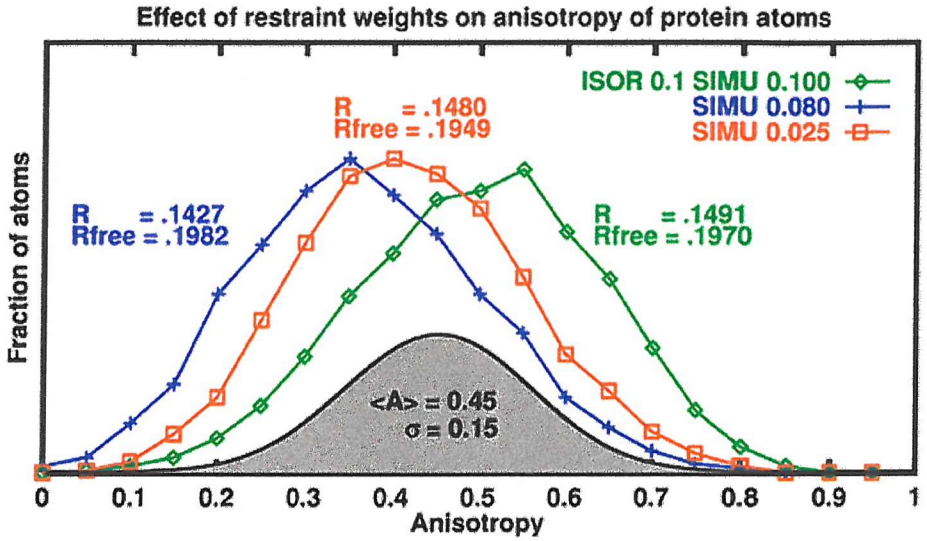


Figure 5.29 Showing the effects on anisotropy caused by alteration of the SIMU restraint in SHELX, taken from Merritt 1999a.

In an extensive study of all protein structures in the protein database refined to a resolution of 1.4 Å or better a global anisotropy value of 0.45 has been calculated (Merritt 1999a). To analyse the calculated ADPs and anisotropies for this structure the pdb file was uploaded to the Parvati server (Merritt 1999a). This calculated a mean anisotropy of 0.73 and accompanying σ of 15 for all protein atoms using the default SIMU value of 0.1. Using the 0.025 SIMU restraint both the R_{factor} and R_{free} dropped by 0.3% with an anisotropy value of 0.765 with a σ of 15.

Endothiapepsin PD-135,040 Complex refinement

This *gem*-diol inhibitor complex was refined in a similar manner as the other high resolution X-ray structures with SHELX. However the diffraction limit of the crystal was around 1.6 Å i.e. around 0.5 Å lower than the data collected from the other crystals. This can be attributed to the fact that the protein crystallised in the weaker diffracting crystal form. Therefore there was not enough data to allow a full anisotropic refinement of all the atoms in the structure. After initial rigid body and isotropic refinements a number of side chains in the structure could be modelled in different conformations these included Ser 36, Ser 74, Thr 88, Ser 109, Thr 116, Thr 127, Gln 134, Ser 145, Asp 140, Ser 145, Thr 164, Ser 204, Val 235, Ser 236, Asp 271, Ser 282, Ser 289 and Val 305. This was done and the occupancies refined for each conformation. The total number of disordered residues is 17 representing 5.15 % of all residues in the structure.

Following this hydrogen atoms were added to the model. The final R_{factor} and R_{free} values from SHELX refinement were 16.38 % and 20.96 % respectively, which are respectable values for a protein studied at a resolution of 1.6 Å. To see if a better model could be obtained refinement was switched to REFMAC (Murshudov *et al* 1997) for its ability to refine TLS parameters with the R_{free} set preserved. The structure was refined until convergence had occurred using maximum likehood targets which gave a final R_{factor} and R_{free} of 15.54 % and 19.11 %. Then the two domains of endothiapepsin (1 to 189 and 190 to 302) where assigned as different TLS groups. Refinement of the TLS groups was done first with all B_{iso} values set to a single value 20. Once the TLS refinement had converged the TLS parameters were then fixed and the atomic co-ordinates and individual B_{iso} values refined for each atom. A number of identical refinements were carried out in which the only difference was in the number of TLS groups. The second set of TLS groups were formed from the α helices and β sheet secondary structure elements of endothiapepsin which gave a total of 15 groups. The residues within these secondary structure groups were defined by use of the program DSSP (Kabsch and Sandler 1983). The use of TLS groups caused little change in the R_{factor} and R_{free}

which were 15.75 % and 19.04 % respectively for the refinement using two domains and 15.76 % and 19.13 % respectively for the TLS refinement using the secondary structure elements. There was also little change in the electron density and B_{iso} values after the TLS refinements. Therefore TLS refinement of this complex was judged to be unjustified and was not included in the final model. The statistics for the final refinement converged with the statistics shown in Table 9 (values for the outer resolution shell are given in brackets). The data completeness in the outer shell is low but as the I/σ(I) value for this shell is above two the data were included in the refinement.

PDB code	Not submitted
Unit Cell	a 53.88 Å, b 73.73 Å, c 45.00 Å, β 110.26 °
Space Group	$P2_1$
Number of unique reflections	47,095
Resolution Range Å	10-1.60
Outer Shell Å	1.69-1.60
Multiplicity	4.4(2.1)
I/σ(I)	5.30 (3.60)
R_{merge}	7.3% (17.4%)
Data completeness	89.0% (52.7%)
R_{factor}	15.54%
R_{free}	19.11%
Number of water molecules	436
Average B_{iso} protein atoms	12.30
Data to parameter ratio	

Table 9 Crystallographic statistics for the endothiapepsin PD-135,040 complex. Figures for the outer shell are given in brackets.

No outliers were detected in the Ramachandran plot of this structure (Figure 5.30).

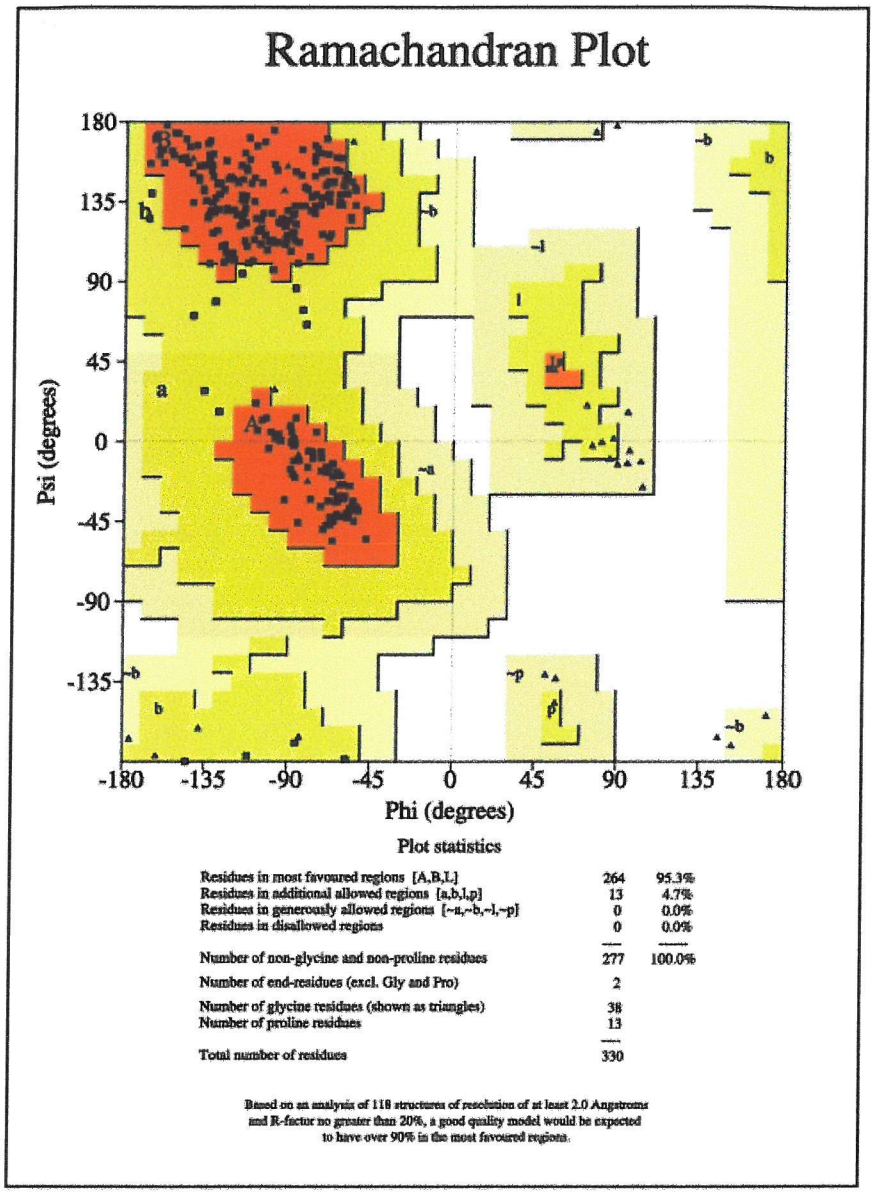


Figure 5.30 A Ramachandran plot of the endothiapepsin PD-135,040 structure after the final refinement.

After the refinement had converged the model was refined against all data including the R_{free} set. After this the ESDs for all the bond lengths in the structure were calculated from an unrestrained least squares matrix inversion. The bond lengths for the catalytic aspartates are shown in Figure 5.31.

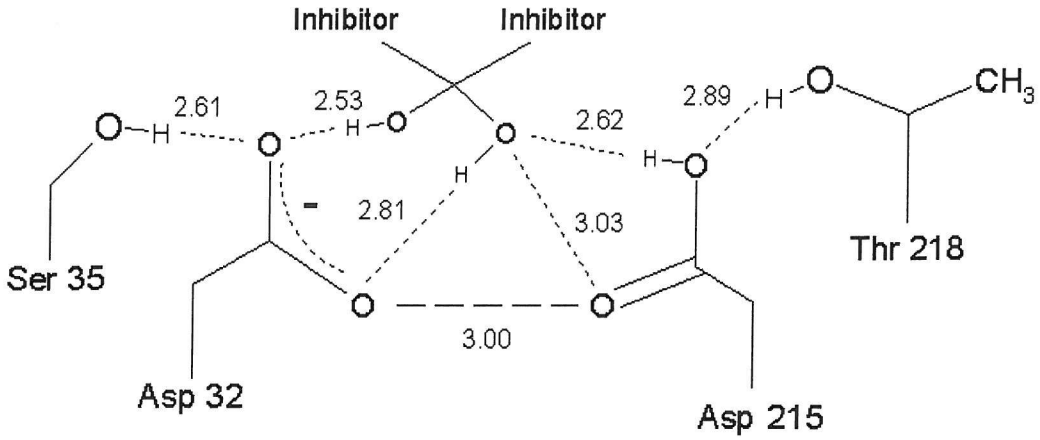


Figure 5.31 Showing the bond lengths in the active site of the PD-135,040 structure. The ESDs for all four carboxyl bond lengths is 0.061 Å.

The electron density around the active while quite acceptable gives no information on the position of hydrogen atoms as might be expected at this resolution (Figure 5.32).

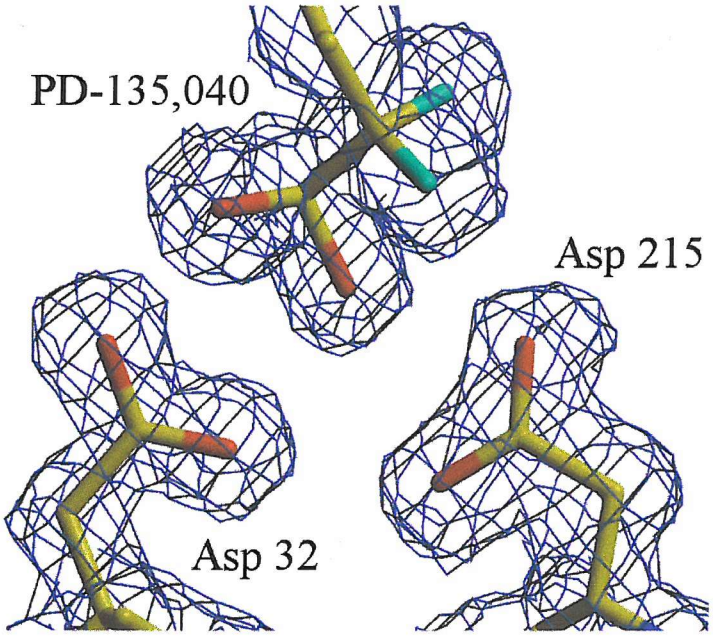
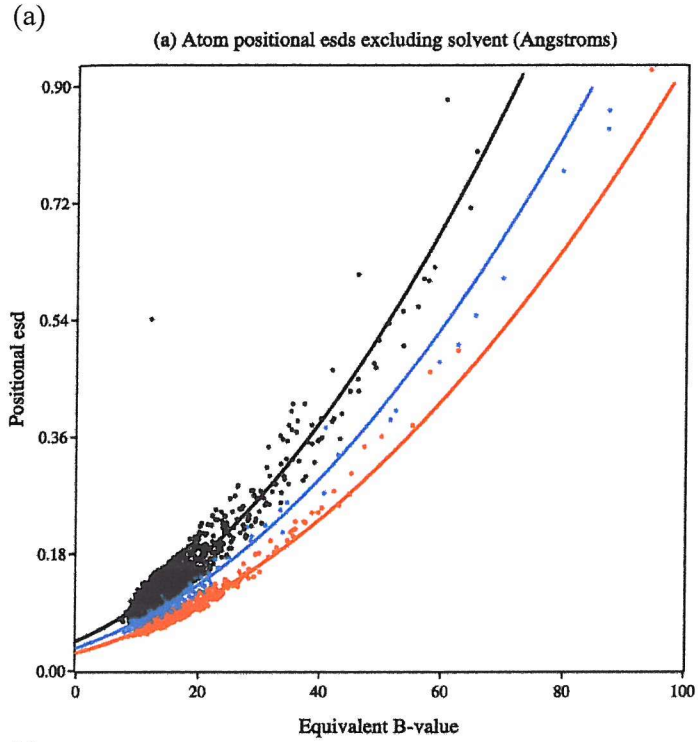


Figure 5.32 Showing the $2mF_o$-DF_c electron density at 1σ around the active site of the PD-135,040 active site. There is no mF_o-DF_c density at $+2\sigma$ or -2σ.

The atom positional and bond length ESD values for all protein atoms are shown in Figure 5.33. The ESD values are higher than those in the other five X-ray diffraction structures as might be expected due to lower resolution of the PD-135,040 structure.

(a)



(b)

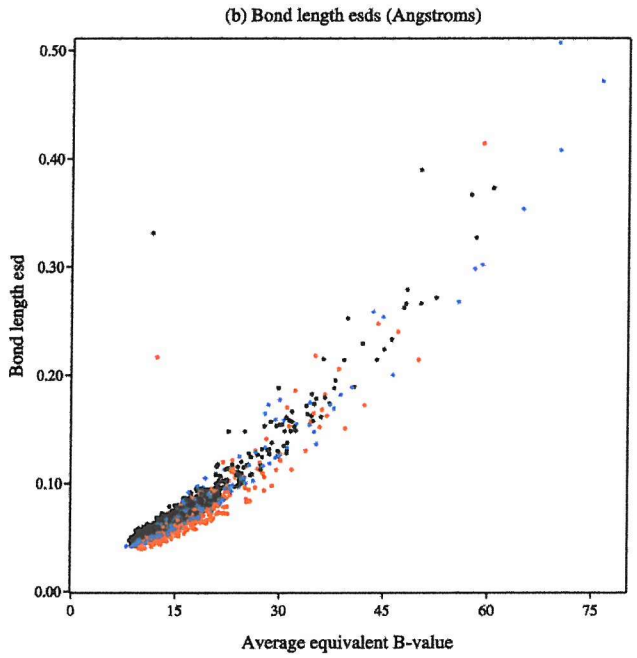


Figure 5.33 The results of an unrestrained least squares matrix inversion for the endothiapepsin PD-135,040 structure. All protein atoms and bond lengths are shown; there are no outliers. In (a) carbon atoms are represented in black nitrogen in blues and oxygen in red while in (b) C-C bonds are represented in black C-N bonds in blue and C–O, C=O bonds in red.

ADP, Anisotropy and ESD analysis of the active site aspartates

The anisotropy of each of the five atomic resolution X-ray structures was determined in distance shells from the centre of mass by use of the program RASTEP (Merritt and Bacon 1997).

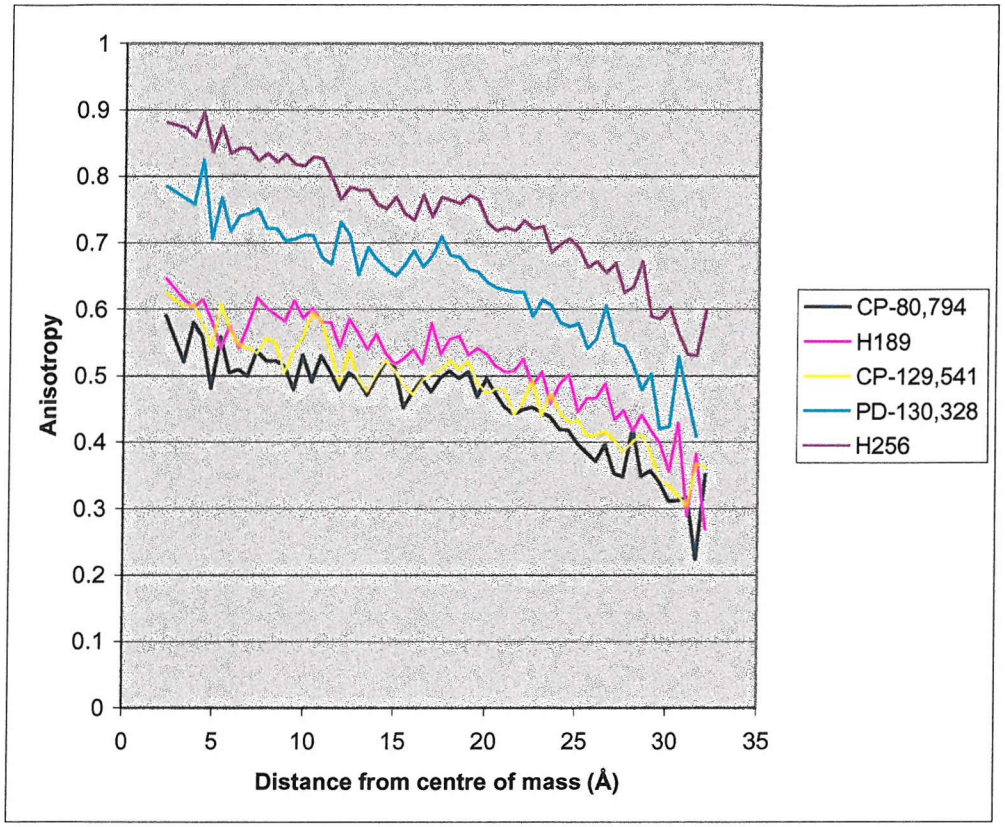


Figure 5.34 Showing the anisotropy in distance shells from the centre of mass for all five atomic resolution structures obtained by X-ray diffraction.

All five of the structures show the same trend in which anisotropy decreases as the distance from the centre of mass increases. This means that atoms close to the centre of mass such as the catalytic aspartates are the most isotropic in the structure with anisotropy decreasing the further away an atom is from the centre of mass. The values for the three statine based inhibitor structures are all clustered together at the bottom of the graph shown in Figure 5.34, while the values for the PD-130,328 and H256 structures show the same trend but with higher anisotropy

values and a steeper curve as the distance from the centre of mass increases. Of the five structures anisotropy is lowest in the CP-80,794 structure. The atomic ADPs reduced to a single B_{iso} value have been averaged on a residue by residue basis for all five atomic resolution X-ray structures and are shown in Figure 5.35. Also shown are the B_{iso} values for the PD-135,040.

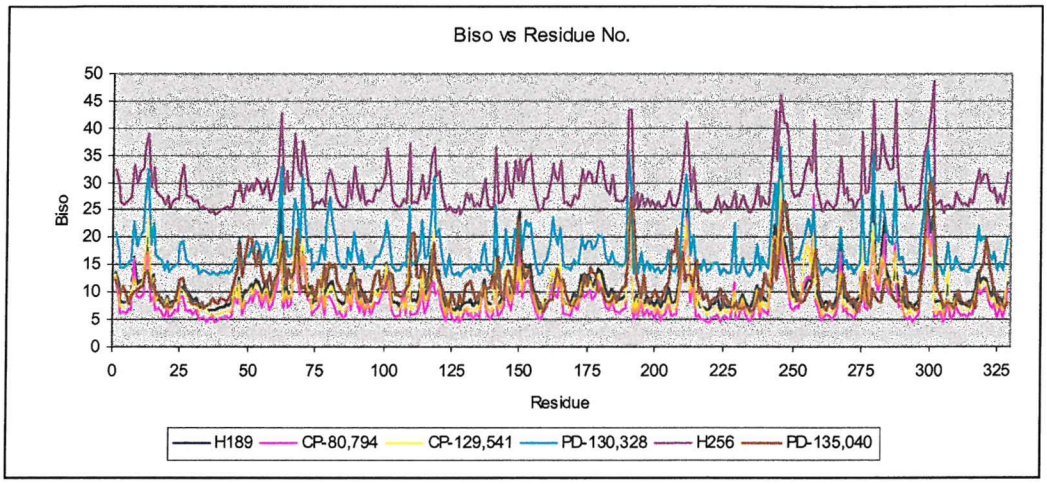


Figure 5.35 Showing the average B_{iso} value for each residue in the H189, PD-130,328, CP-80,794, H256, PD-129,541 and PD-135,040 structures.

The B factor plots for all six structures have the same overall pattern with higher average B_{iso} values for the PD-130,328 and H256 structures compared to the three statine inhibitor based structures. The peaks in the B_{iso} plot are associated with loops within the protein structure as might be expected. The lowest B_{iso} values can be found in the CP-80,794 structure. The B_{iso} values for the PD-135,040 structure which was refined isotropically compare well with the values for the statine based inhibitors.

Electron Density

The electron density is well defined in all six X-ray diffraction structures. However the electron density in the statine based inhibitor structures was extremely well defined around the atoms in each structure. The electron density in the H256 and PD-130,328 structures while good, lacked the degree of atomicity present in the statine based inhibitor structures. Perhaps the most atomistic electron density is present in the CP-80,794 structure which is probably due to the extremely low B_{iso} values for the atoms in this structure. In this structure electron density is visible for a number of hydrogens, the hydrogen in the hydrogen bond between Thr 88 and Asp 87 is visible. The length of this hydrogen bond is 2.61 Å and an atom modelled into the difference density between the two residues is 1.21 Å from the OG1 atom on Thr 88 and 1.40 Å from the OD1 atom of Asp 87. The C-O-H bond angle for Thr 88 is 103.80° and 149.50° for Asp 87. All of which would seem to indicate that the hydrogen atom is more closely associated with Thr 88 (Figure 5.36). This provides further evidence that Asp 87 is negatively charged at the crystallisation pH of 4.5.

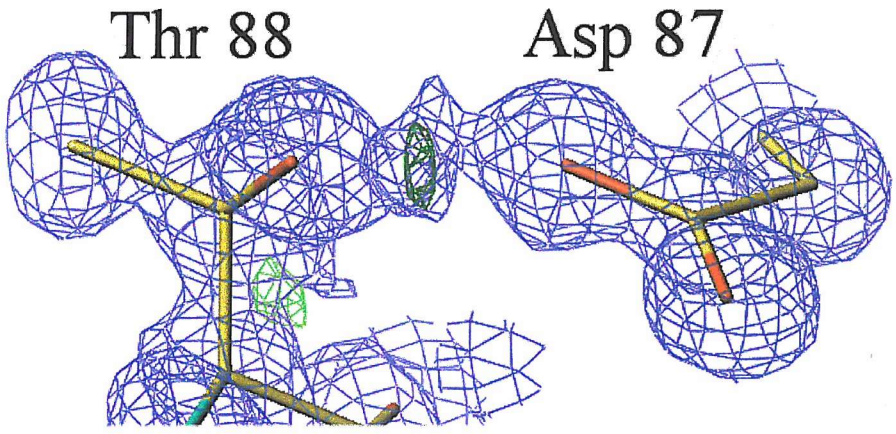


Figure 5.36 The electron density around Asp 87 and Thr 88. The $m2F_o$-DF_c at 1σ is shown in blue and the mF_o-DF_c density at +3 σ is shown in green and at -3σ in red.

Protein Aging

In all of the structures determined, except. for the PD-135,040, Asp 51 and Gly 52 have cyclised to form a succinimide (Figure 5.37) which is a well known characteristic of protein aging (Stephenson and Clarke). In all structures the $2mF_o$-DF_c density is very well defined around the succinimide and adjacent residues indicating that cleavage of the main chain has not taken place.

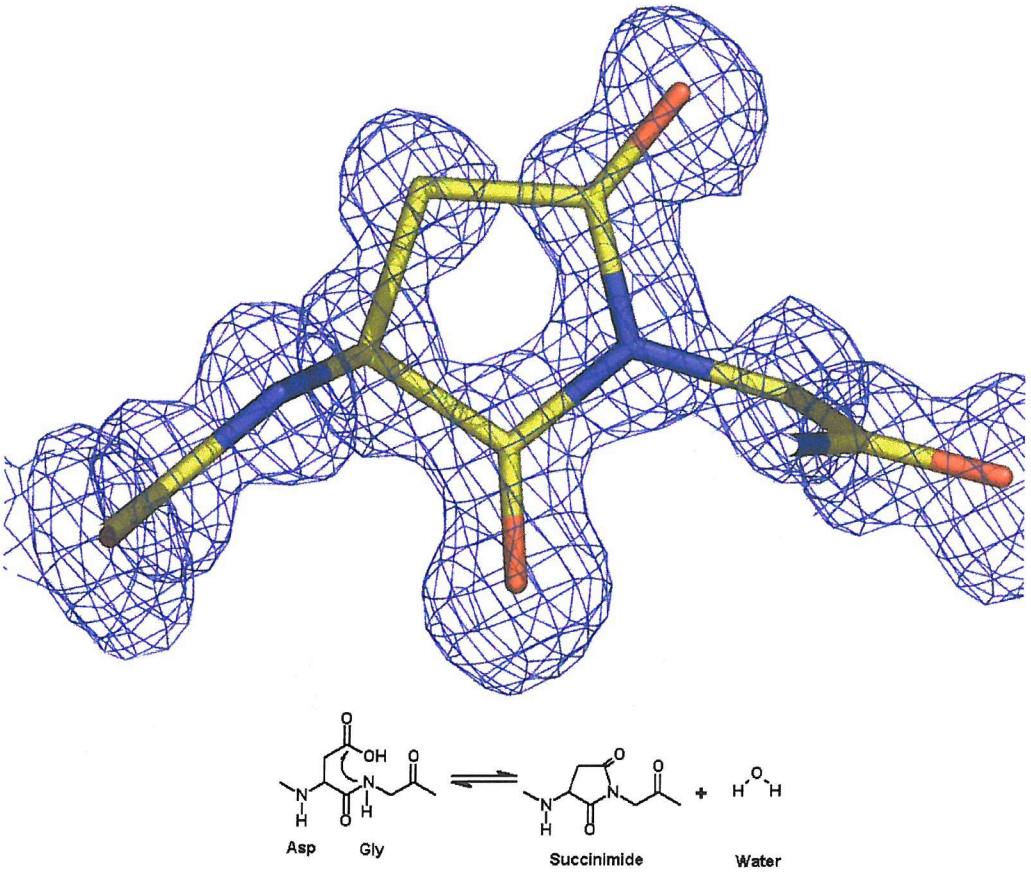


Figure 5.37 The 1σ $2mF_o$-DF_c density for residues Asp 54 and Gly 55 which have cyclised to form a succinimide. Superimposed is an outline of the mechanism of succinimide formation, the carboxyl group of the aspartate is attacked by the nitrogen of the following glycine residue. The succimimide is then formed by intermolecular cyclisation.

Multiple conformations

A number of side chains in the six X-ray structures were modelled and refined as multiple conformations. Several of these were serine residues on the outside of the protein. There were no multiple conformations within the hydrophobic core of the protein in any of the structures. After the introduction of a multiple conformation the electron density and refined side chain occupancies were checked after refinement. The presence of 1σ $2mF_o$-DF_c density for the multiple conformation and removal of 2.0 σ mF_o-DF_c density coupled with an occupancy above 0.1 was used to signify a valid conformation. An example of a multiple conformation is shown in Figure 5.37.

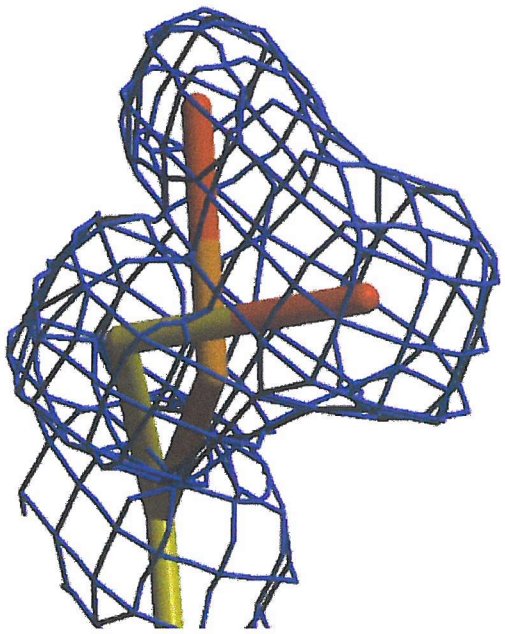


Figure 5.38 The two conformations of Ser 39 as modelled into the CP-80,794 structure. The occupancy for the yellow sidechain is 0.6 and 0.4 for orange sidechain. Each on these conformations is within hydrogen bonding distance of a water molecule.

The protonation states of the non-catalytic aspartates

After the unrestrained positional refinement of all the carboxyl oxygens in each structure the difference in the carboxyl bond lengths were calculated to deduce the protonation state (Deacon *et al* 1997). The ESDs for the difference between the carboxyl C-O bond lengths (σ_D) was calculated using the formula below and then compared with the carboxyl bond length difference using the formula below.

$$\sigma_D = \sqrt{\sigma_1^2 + \sigma_2^2}$$

In which σ_1 and σ_2 represent the ESDs in the carboxyl bond lengths for each aspartate. Those aspartates with a bond length difference greater than two and a half times the ESD (σ_D) were designated as protonated while those carboxyls with bond length differences less than two and a half times the ESD variance were designated as negatively charged. This technique was used to analyse the aspartates in different structures to see if a conserved protonation state could be determined for the carboxyl groups in the structure. It has been speculated that some of the buried carboxyl side chains remain unprotonated at low pH and do not interact in salt bridges. Aspartate 87 has been identified in the neutron diffraction study as a buried negatively charged residue (Coates *et al* 2001). This residue is almost completely conserved in the aspartate proteinases. We aimed to determine the protonation state of this aspartate in the atomic resolution structures to check the assignment made from the neutron structure and to try and identify any more possible buried negatively charged aspartates.

The deduced protonation states for all of the aspartates in each of the statine based inhibitor structures are given in Table 10. The H256 and PD-130,328 structures have much higher B_{iso} values than those of the statine based inhibitors and since the statine based inhibitor structures have reduced displacement parameters they give more accurate atomic positions (Dauter *et al* 1997). Apparent bond lengths can be influenced by the B_{iso} values of the atoms involved and so the H256 and PD-130,328 structures where omitted from this analysis. The high B_{iso} values for

these structures could be attributed to the larger size of these crystals compared to the crystals for the other inhibitors. Larger crystals can give rise to higher mosaic spreads on cooling to cryogenic temperatures (Garman 1999) thus increasing apparent atomic displacement.

Aspartate Residue	H189 Assignment	CP-80,794 Assignment	CP-129,541 Assignment	State in all Structures	Mean ESD Å	Location in protein
8	Charged	Charged	Charged	**Charged**	0.017	Surface
11	Protonated	Protonated	Protonated	**Protonated**	0.015	Surface
12	Protonated	Charged	Charged		0.019	Surface
30	Charged	Protonated	Protonated		0.011	Buried
32	Protonated	Protonated	Protonated	**Protonated**	0.010	Buried
37	Protonated	Charged	Protonated		0.011	Buried
77	Charged	Charged	Charged	**Charged**	0.015	Surface
83	Protonated	Charged	Charged		0.016	Surface
87	Charged	Charged	Charged	**Charged**	0.013	Buried
114	Split	Split	Split	Split	Split	Surface
118	Charged	Charged	Charged	**Charged**	0.011	Surface
140	Charged	Charged	Charged	**Charged**	0.024	Surface
147	Charged	Charged	Charged	**Charged**	0.022	Surface
154	Charged	Charged	Charged	**Charged**	0.014	Surface
171	Charged	Charged	Charged	**Charged**	0.015	Surface
211	Charged	Protonated	Charged		0.022	Buried
215	Protonated	Protonated	Protonated	**Protonated**	0.011	Buried
271	Charged	Charged	Charged	**Charged**	0.048	Surface
274	Charged	Charged	Charged	**Charged**	0.031	Surface
304	Charged	Charged	Protonated		0.041	Buried
Glutamate Residue	**H189 Assignment**	**CP-80,794 Assignment**	**CP-129,541 Assignment**	**State in all Structures**	**Mean ESD Å**	**Location in protein**
44	Split	Split	Split	Split	Split	Surface
49	Protonated	Protonated	Protonated	**Protonated**	0.013	Surface
102	Charged	Protonated	Protonated		0.011	Buried
113	Protonated	Charged	Charged		0.022	Surface
191	Charged	Protonated	Charged		0.015	Surface

Table 10 Showing the deduced protonation states of the all the aspartate and glutamate residues in a single conformation in the H189, CP-80,794 and CP-129,541 structures.

The ESD values of the carboxyl bond lengths in the catalytic aspartates are around 0.010 Å for Asp 32 and 0.011 Å for Asp 215 for all three structures and represent the most ordered aspartates in each of the three statine inhibitor based structures. The ESDs of the C-O bond lengths of the catalytic aspartates are constantly lower than the average aspartate C-O bond ESD. In the H189 structure the average aspartate C-O bond length ESD is 0.0173 Å whereas the ESDs of the catalytic aspartates are at least 40 % lower ranging at around 0.01 Å. In the CP-80,794 structure the average aspartate C-O bond ESD is 0.0165 Å while the maximum ESD in the active site (0.011 Å) is 35 % lower than the average value. The low ADP and ESD values for the atoms in the active site are probably caused by the network of supporting hydrogen bonds from surrounding residues, which act to stabilise the conformations of the catalytic residues. Interestingly Asp 32 and Asp 215 appear to both have C-O bond lengths characteristic of protonated carboxyl groups in all three structures. This is not likely due to chemical considerations indicating the possibility of a more complex arrangement in the active site.

Of the 19 aspartates in a single conformation, the protonation states of 13 are conserved within all three of the structures. However these assignments are better for some aspartates than others. Asp 87 is negatively charged in all three structures with a mean carboxyl bond length ESD of 0.0125 Å confirming the assignment made by Coates *et al* (2001) in a neutron diffraction study of endothiapepsin complexed with a hydroxyethylene based inhibitor H261. Aspartates 118 and 154 also appear to be negatively charged in all three structures and have similar ESD values to Asp 87 although they are located on the surface of the protein. The protonation states of aspartates 271, 274 and 304 are less certain than that of the other aspartates as the bond length ESD values for these residues are somewhat higher at around 0.048 Å, 0.0308 Å and 0.041 Å respectively in all three structures. Accordingly these aspartates are located in a more mobile part of the protein. Of the four glutamate residues in a single conformation only Glu 49 has the same protonation state (protonated) in all three statine structures.

Electrostatic potential

The electrostatic potential of endothiapepsin was calculated using GRASP (Nicholls *et al* 1991) with and without the information on the protonation states of the aspartate and glutamate residues. All of the preserved protonation states in the three statine inhibitor based structures (Table 10) were entered into one model. For aspartates and glutamates where all three protonation states were not the same the consensus protonation state was entered (two out of three). This generated an 'actual' electrostatic potential at the crystallisation pH of 4.5 and a default potential for endothiapepsin at pH 7.0 (Figure 5.39).

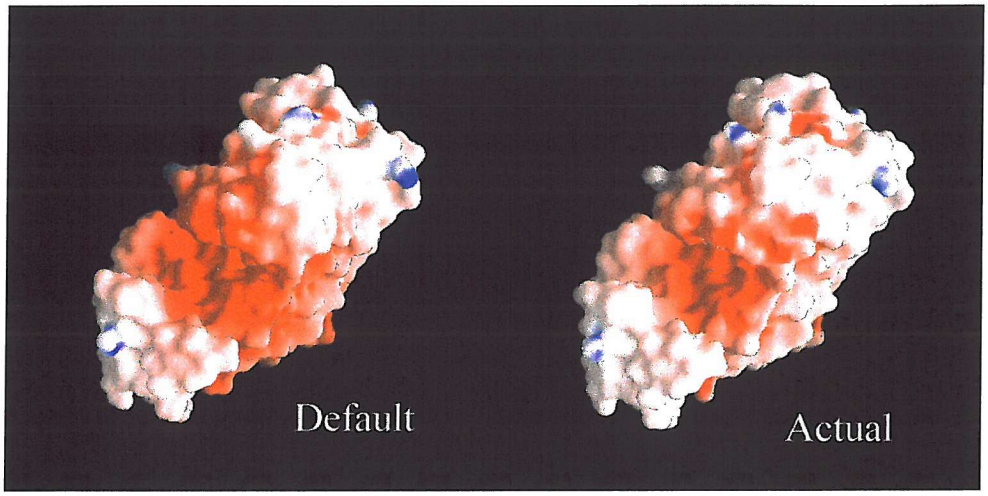


Figure 5.39 Showing two electrostatic potentials of endothiapepsin at pH 4.5. The model produced using the default charge settings is more negatively charged around the substrate binding site. The actual model uses the derived protonation states is much less negatively charged.

The model produced using the default GRASP charge settings (pH 7.0) in which all aspartate carboxyl oxygens have a charge of -0.5 is more negatively charged around the substrate binding site. The introduction of protonated aspartates with a charge of 0 into the actual model reduces the negative electrostatic potential around the substrate binding site which more accurately models the electrostatic potential at the crystallisation pH 4.5.

1D H NMR Spectra

To help verify the very short hydrogen bonds in the active site a number of 1D H NMR spectra where run on native endothiapepsin and endothiapepsin complexed with a series of inhibitors (Figure 5.40).

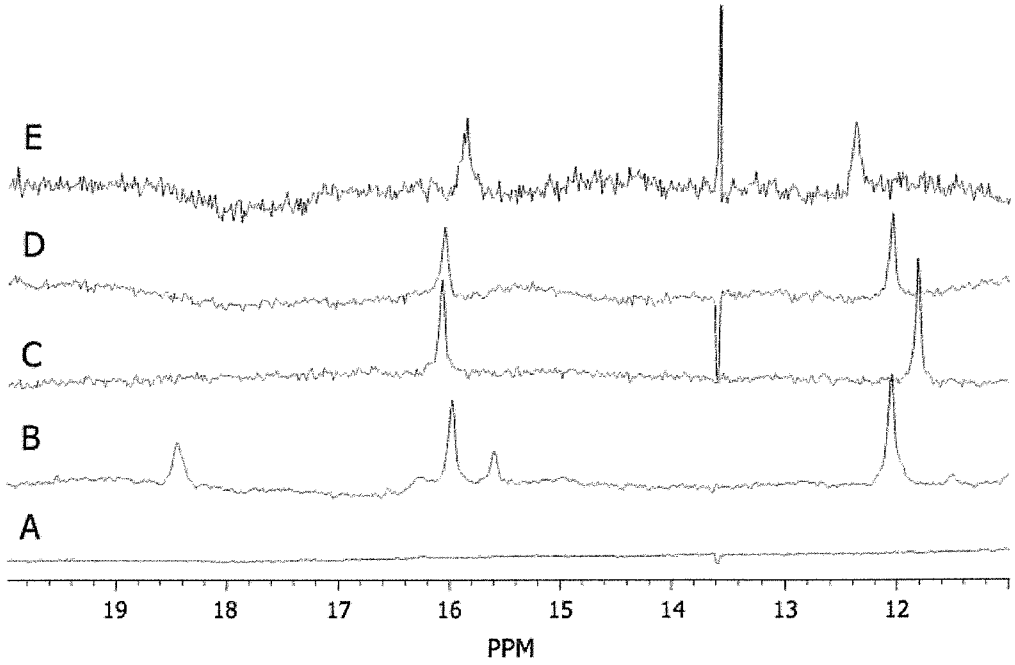


Figure 5.40 1D H NMR spectra of free endothiapepsin (a), endothiapepsin with a phosphinic acid analogue inhibitor (b), endothiapepsin with PD-135,040 a *gem*-diol inhibitor (c), endothiapepsin with a statine based inhibitor (d) and endothiapepsin with a reduced bond inhibitor (e). These spectra where collected with the protein in 90 % H_20 / 10 % D_2O at a temperature of 283.1 K.

As LBHBs have proton NMR chemical shifts far downfield of normal protein signals typically between 16-21 ppm, a number of 1D H NMR spectra were run to confirm the presence of the LBHBs suggested by the crystal structures. These spectra where collected with the protein in 90 % H_20/10 % D_2O at a temperature of 283.1 K and are shown in Figure 5.40. The NMR spectrum of free endothiapepsin shows no peaks outside the normal region for protein signals 1-11 ppm. However when complexed with an inhibitor peaks are visible at around 16 and 18 ppm. In the endothiapepsin statine inhibitor spectrum there is a sharp peak

at 16.1 ppm consistent with the short hydrogen bond found in the statine structures between the carboxyl oxygen of Asp 32 and the inhibitory statine hydroxyl. The length of this hydrogen bond length ~2.6 Å is consistent with the peak at 16.1 ppm (McDermott & Ridenour 1996). The spectra thus indicate the presence of LBHBs in the endothiapepsin complex at 283 K close to physiological temperatures. In the spectra of endothiapepsin complexed with a phosphinic acid analogue inhibitor the peak at 16.1 ppm is supplemented by a peak at 18.7 ppm. This peak corresponds to a very short hydrogen bond between the outer oxygen of Asp 32 and the O2 atom on the phosphinate group (2.41 Å). Another short hydrogen bond exists between the outer oxygen of Asp 215 and the O1 atom in the phosphinate group in the PD-130,328 structure (2.57 Å) probably corresponding to the peak at 16.1 ppm. In the spectrum of the *gem*-diol inhibitor PD-135,040 complexed with endothiapepsin there is a single broad peak at around 16.1 ppm which relates to a hydrogen bond length of ~2.60 Å. In the 1.6 Å structure of PD-135,040 the hydrogen bond length between the outer oxygen of Asp 215 and an inhibitor hydroxyl is 2.62 Å. The other shorter hydrogen bond (2.53 Å) between OD2 of Asp 32 and the other inhibitor hydroxyl is probably represented in the same peak. This bond length was found to be 2.58 Å in the previous 2.30 Å X-ray structure (PDB code 1epr).

Chapter 6

Discussion

A number of different techniques have been used to identify the protonation states of the two aspartates when a transition state analogue inhibitor is bound to the active site of endothiapepsin. All of the work conducted indicates that Asp 32 is the negatively charged aspartate when a transition state analogue inhibitor binds to the enzyme. The $2mF_o$-DF_c density map of the active site from model 1 of the H261 neutron diffraction structure (Figure 5.07) clearly shows that outer oxygen of Asp 215 is deuterated. This model is also consistent with the density present in the unbiased difference map of the active site. Evidence for the presence of a deuterium on Asp 215 can also be seen in the Fourier maps for model 2 (Figure 5.08) in which a patch of positive density was found close to the outer oxygen of Asp 215 indicating a deuterium is attached to the outer oxygen of Asp 215. Occupancy refinement of the protons within the two models also indicates that model 1 is the most likely model of the active site as the occupancy of both deuterium atoms is close to unity (Figure 5.09). However the fact that the neutron data do not resolve completely between models 1 and 2 may indicate some extra properties of the catalytic mechanism. The donor acceptor distances for the hydrogen bonds made between the inhibitor hydroxyl and the catalytic aspartates are very short around 2.6 Å (Figure 5.41), this distance is confirmed in all of the X-ray studies except for H256, which is the weakest inhibitor. The hydrogen bond between the inner oxygen of Asp 32 and the inhibitory group (~2.6 Å) is consistently the shortest in both the neutron and atomic resolution X-ray structures of the statine based inhibitors (H189, PD-129,541 and CP-80,794). In the structure of the *gem*-diol PD-135,040 complexed with endothiapepsin this short hydrogen bond is formed between the outer oxygen of Asp 32 and one of the inhibitor hydroxyls (Figure 5.31). This situation is the same in the PD-130,328 structure (Figure 5.13). The short hydrogen bond between the second inhibitory group and Asp 215 is formed with the inner oxygen in the PD-130,328 structure and the outer oxygen in the PD-135,040 structure.

Some of the difficulty in discriminating between the two models in the neutron structure could be due to the formation of a LBHB between the hydroxyl of the

inhibitor and the inner oxygen of Asp 32. Whilst the neutron results indicate that Asp 215 is protonated in the transition state complex, the formation of a LBHB in the active site between the inner oxygen of Asp 32 and the inhibitory statine hydroxyl would tend to delocalise neutron density for the deuteron between these two groups. As hydrogen/deuterium atoms involved in LBHBs will be distributed between two positions around 0.5 Å apart (Cleland *et al* 1998) and may therefore have high apparent temperature factors making them difficult to locate. This delocalisation of density can be seen in the neutron structure between Asp 32 and the statine hydroxyl. The active site bond lengths for all of the statine based inhibitors (H189, PD-129,541, CP-80,794) are all very similar (Figure 5.41).

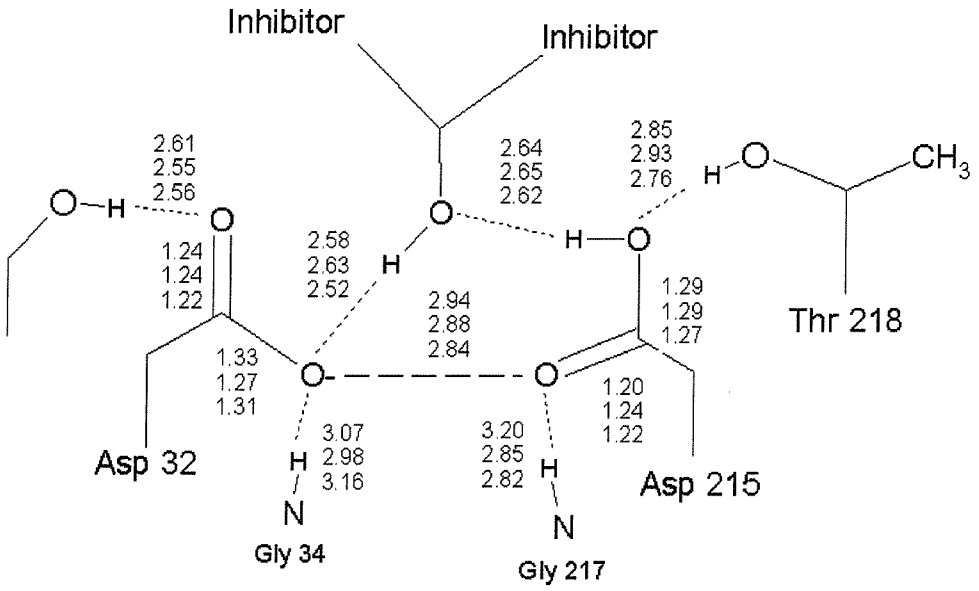


Figure 5.41 The bond lengths at the active site for the three statine based inhibitors H189, CP-80,794 and CP-129,541 are shown in descending order in (Å). Hydrogen bonds are indicated by dotted lines. The ESD values for the carboxyl bonds range between 0.009 Å and 0.013 Å.

The shortest hydrogen bond in all three of these structures occurs between the inner oxygen of Asp 32 and the statine hydroxyl oxygen. The length of this bond is 2.58 Å in the H189 complex, 2.63 Å in the CP-80,794 structure and 2.52 Å in the CP-129,541 structure. The short length of this bond in all three structures combined with the exclusion of any solvent from the active site indicates that this

bond is likely to be a LBHB as indicated by the neutron diffraction study of H261. Further evidence for the presence of a LBHB between the inner oxygen of Asp 32 and statine comes from the ADPs of the catalytic aspartates. In all of the X-ray structures examined the ADPs are lower for Asp 32 indicating better ordering in the atoms in Asp 32, which could be the result of a LBHB. The electron density in the active site of the CP-80,794 structure can also be explained by this model. There is also excellent positive $2mF_o$-DF_c and mF_o-DF_c electron density in the CP-80,794 model for the presence of a hydrogen on the statine hydroxyl orientated towards the inner oxygen of Asp 32 with a C-O-H bond angle of 109.05° and an O-H bond length of 1.24 Å (Figure 5.21).

This indicates that inner oxygen of Asp 32 is unlikely to be protonated when a transition state analogue inhibitor is bound, which is in agreement with the mechanism proposed by Veerapandian (Figure 1.03) but is in conflict with other proposed mechanisms such as that outlined in Suguna *et al* (1987). The fact that this proton cannot be seen in the other two statine inhibitor complexes could be due to the fact that the distance between the inner oxygen of aspartate 32 and the oxygen on the statine hydroxyl (2.63 Å) is slightly longer than the same bond in the other two structures in which formation of a stronger LBHB could reduce the electron density for the hydrogen on the statine hydroxyl.

The aspartate C-O bond lengths present in all of the statine like inhibitors would seem to indicate the protonation of both aspartates. This is not likely because of geometric considerations in the active site, neither is it likely that the inhibitory statine oxygen is not protonated as the pK_a of a hydroxyl group is 6 pH units higher than that of an aspartate. This would seem to indicate that one of the aspartates is ionised and the negative charge on the charged aspartate is not shared equally between its two C-O bonds but is localised to a single C-O bond. The bond length differences on both aspartates could make discrimination between the protonated aspartate and the locally charged aspartate difficult. However a local negative charge on the inner oxygen of Asp 32 would be in keeping with the idea of a LBHB between Asp 32 and the statine hydroxyl. The stabilisation of this

group would be aided by the short hydrogen bond between the statine hydroxyl and the inner oxygen of Asp 32 and the short hydrogen bond between Ser 35 and the outer oxygen of Asp 32. The weak hydrogen bond involving the inner oxygen of Asp 32 and N-H group of Glycine 34 may also help stabilise the negative charge. This model would account for the bond lengths found in the X-ray structures of the statine like inhibitors. It would also explain the low occupancy for a proton on Asp 32 in the neutron diffraction model. And also explain the results of earlier studies in which mutation of Ser 35 to Ala causes a decrease in catalytic activity, Ser 35 unlike Thr 218 is conserved in all aspartic proteinase structures. Threonine 218 is replaced by a serine on a number of aspartic proteinase structures. This conservation of Ser 35 coupled with extremely low ESDs for this residue would seem to indicate a role in catalysis such as the stabilisation of a negative charge on Asp 32. The short hydrogen bond from Ser 35 to the outer oxygen of Asp 32 would also help to form a rigid active site, mutation of this residue in pepsin and other aspartic proteinases causes a 10 fold drop in k_{cat} but causes little change in pH optimum (Lin *et al* 1992) suggesting that Ser 35 and its counterpart Thr 218 are required to maintain the structure of the active site.

The atomic X-ray diffraction of the PD-130,328 acid complex backs up model 1 from the neutron experiment. The C-O bond lengths of the catalytic aspartates from unrestrained refinement clearly indicate that Asp 32 with its similar C-O bond lengths is negatively charged and unprotonated (Figure 5.13). The different C-O bond lengths in Asp 215 confirm it as the protonated aspartate when PD-130,328 is bound. The shorter bond between the γ carbon and the outer oxygen (1.22 Å) indicates the presence of a double bond between the two atoms. While the longer bond between gamma carbon and inner oxygen (1.33 Å) indicates that there is a proton attached to the inner oxygen atom. The fact that a proton is attached to the inner oxygen of Asp 215 and not the outer oxygen as would be expected in the Veerapandian mechanism could be explained by the distortion of the active site induced by the binding of the large phosphate group to the active site. This distortion is also likely to be partly responsible for the high K_i value of the inhibitor (110 nM). The LBHB (2.41 Å) between the outer oxygen of Asp 32

and O2 of the phosphinate group is the shortest hydrogen bond in all the structures determined. Bonds of this length have been observed between aspartates and phosphate groups previously (Wang *et al* 1997) in which a LBHB of 2.43 Å was observed in the high resolution structures of a protein receptor phosphate complex between OD2 of asp 56 and O4 of a phosphate group. The electron density for these atoms "was nearly touching" which is certainly the case in the PD-130,328 structure where there is a minimal gap in electron density between O2 on the phosphate and OD2 of Asp 32, a feature not seen for other hydrogen bonds involving these inhibitors.

The hydrogen bonding pattern in the active site of the *gem*-diol structure (PD-135,040) also suggests that Asp 32 is negatively charged as both the carboxyl oxygens are forming hydrogen bonds with the inhibitor hydroxyls. The outer oxygen atom of Asp 32 would appear to be involved in two short hydrogen bonds, one with the hydroxyl of Ser 35 and one with the outer hydroxyl group of the inhibitor (Figure 5.31). This would suggest that when the *gem*-diol inhibitor is bound the LBHB network involves the outer oxygen of Asp 32 not the inner oxygen as observed in the single hydroxyl statine based structures.

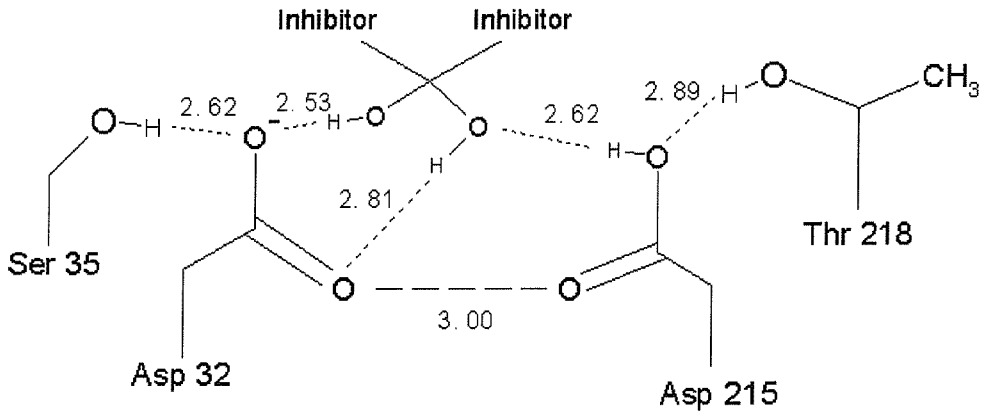


Figure 5.42 Showing the possible bond arrangements and hydrogen positions within the active site of the PD-135,040 structure.

In Figure 5.42 the outer oxygen of asp 32 is involved in two LBHBs one to Ser 35 and another to an inhibitory hydroxyl both of which are likely to be protonated at

the crystallisation pH of 4.5. This arrangement could be used to stabilise a negative charge on the outer oxygen of asp 32 which could be present in the transition state. The resolution of the PD-135,040 structure is 1.6 Å and the ESD values on the carboxyl bonds of the catalytic aspartates are around 0.06 Å, which is too large to allow differentiation between single and double bonds. The model shown in Figure 5.42 is thus the most likely based on the neutron and atomic resolution X-ray structures. The structure does confirm the presence of LBHBs (~2.6 Å) in the active site of the *gem*-diol endothiapepsin complex which can be matched to a peak in the NMR spectra of this complex. Kinetic studies on HIV proteinase using ^{15}N labelled substrates in H_2O and D_2O indicated the protonation of the nitrogen atom in the substrate peptide bond during catalysis (Rodriguez *et al* 1993). This study also predicated a catalytic mechanism in which a LBHB is formed between the outer oxygen of Asp 25 and one of the inhibitory hydroxyls in the transition state and the protonation of Asp 25' on the outer oxygen. This model is consistent with the crystallographic and kinetic data on the enzyme and is a close match to the endothiapepsin *gem*-diol transition state analogue model. For a number of years LBHBs have been implicated in the catalytic mechanism of serine proteinases it would now seem that they are involved in the aspartic proteinases.

To sum up, the 2.1 Å neutron diffraction structure of the endothiapepsin H261 complex clearly indicates that the outer oxygen of Asp 215 is protonated when the transition state analogue inhibitor is bound at the active site (Figure 5.07). The neutron diffraction model does not clearly resolve the protonation state of Asp 32. However the electron density in the endothiapepsin CP-80,794 structure indicates that the hydrogen in the inhibitory hydroxyl is oriented towards the inner oxygen of Asp 32 (Figure 5.21) meaning that Asp 32 is likely to be negatively charged as a proton on the outer oxygen of Asp 32 cannot form a hydrogen bond to the inhibitory hydroxyl. All of which is in agreement with the mechanism outlined in Veerapandian *et al* 1992. The results of the atomic resolution X-ray diffraction bond length analysis of the catalytic aspartates are less well defined. The consensus view from the three statine based inhibitor structures is that Asp 32 and Asp 215 have carboxyl bond lengths which suggest that both are protonated which

is unlikely as the active site has been shown to carry a charge of -1 (Fruton *et al* 1976). However these results could be explained by the stabilization of one of the canonical forms of a negatively charged Asp 32 in which the inner oxygen is negatively charged.

Future work on endothiapepsin will involve the perdeuteration of the enzyme in which all hydrogen atoms will be replaced with deuterium via the recombinant expression of the protein in E. coli or yeast and growth in deuterated media. This should enable higher resolution data to be collected due to the higher signal to noise ratio caused by the removal of incoherent scattering from hydrogen atoms in the protein crystal. This should be done on various endothiapepsin inhibitor complexes and would be feasible with smaller protein crystals due to the decreased background and improved structure factors. However the low flux of neutrons currently available will impose a limit on the resolution to which structures can be determined. Improved neutron sources with much higher fluxes such as the European spallation source with much higher neutron flux would be able to produce higher resolution and thus more accurate neutron structures which would be a major advantage in experimentally locating protons in protein crystals. Further NMR experiments should also be carried out to determine the movement of the two domains on binding various transition state inhibitors.

Higher resolution X-ray diffraction data could also be collected to allow more meaningful analysis of the aspartate carboxyl bond lengths if well ordered enough crystals could be grown. Care would be needed in collected such high resolution data as only a small number of beam lines in the world can collect data to such high to such high resolution. X-ray diffraction data has been collected on aldose reductase to 0.66 Å which has enabled the valance states of most of the atoms in the structure to be determined and the proposal of a new catalytic mechanism for the enzyme. However even at this resolution only 54% of the hydrogen atoms are visible in electron density, there is no electron density for the hydrogen atoms attached to more mobile atoms in the structure. This is a major factor limiting the location of hydrogen atoms using atomic resolution X-ray diffraction.

There is a clear need for a good knowledge of the aspartic proteinase mechanism to help develop drugs to treat serious medical conditions such as Alzheimers disease and malaria. Treatments to both of these diseases could potentially be developed by inhibiting the aspartic proteinases involved in each respective disease.

Appendix 1

Advanced crystallography

Data Integration

The intensity of a reflection or Bragg peak on a diffraction pattern (I) is the defined as the total intensity I_{tot} minus the background intensity I_{bkg} thus

$I=I_{tot}-I_{bkg}$

However it is impossible to measure the background under a diffraction spot. Thus the background is measured in a region around the spot in two dimensions X and Y. A background plane is then fitted to these background pixels and this can then be used to estimate the background underneath the spot. To do this a pixel mask or raster is centred on the predicted spot position and used to calculate the level of the background plane. The pixel mask contains the dimensions of the average Bragg peak as well as the dimensions of the non peak background of the spot. This combined with the estimated mosaic spread is used to measure the intensity of the Bragg peak. It is critical that the mosaic spread value is realistic since if it is not well defined, then either the total intensity of the peak will not be sampled or too large an area of background will be sampled reducing the $I/\sigma(I)$ value. The error associated with a given reflection is given by it's $\sigma(I)$ value which can be worked out from Poisson statistics as

$\sigma^2(I)=I_{tot}+I_{bkg}$

So

$$\frac{I}{\sigma(I)}=\frac{I_{tot}-I_{bkg}}{\sqrt{I_{tot}+I_{bkg}}}$$

Thus when $I/\sigma(I)$ equals one the error in intensity associated with the reflection is equal to the intensity of the peak. Thus resolution shells which have an average $I/\sigma(I)$ much less than 2 should not be used for data refinement as below this ratio the intensity and the error associated with it become equal. After determining the background plane the intensity of each spot is derived using profile fitting, which can typically provide a reduction in variance of two for weak reflections compared to summation integration. In this procedure it is assumed that the shape or profile of the spot is known in two or three dimensions with the intensity being derived by determining the scale factor. Which when applied to the known spot profile gives the best fit to the observed spot profile (Leslie 2000). The scale factor is proportional to the profile fitted intensity for that reflection, this fitting is done by least squares methods to minimise the residual R.

$$R = \sum w_i (X_i - KP_i)^2$$

Where X_i is the background subtracted intensity at pixel i, P_i is the value of the standard profile at the corresponding pixel, w_i a weight derived from the expected variance of X_i, and K the scale factor to be determined. This procedure does assume a knowledge of the true reflection profile which is determined from the observed reflection profiles of a number of reflections in the immediate vicinity of the reflection being integrated. An approximate weighted sum of the individual profiles is used to form the true or standard profile. The shape of a reflection will vary with its position on the detector due to changes in the obliquity of incidence and this must be accounted for. The most useful values given in the MOSFLM summary file are the residual (R) and the weighted residual (WRESID). The residual is the rms positional residual in mm after refinement of the detector parameters. To calculate this MOSFLM predicts where a spot should ideally be and measures the distance between this and actual spot location. An rms distance deviation is calculated for the whole image. Values of between 0.02-0.04 are ideal for images recorded from CCD detectors with their small pixel size and indicate that unit cell parameters are correct. The WRESID should be close to unity

indicating the residual is independent of the strength of the diffraction image. After checking the summary file for errors the next step is to sort the MTZ file using SORTMTZ using the hkl indices and the intensity for each reflection. After the MTZ file has been sorted it can be merged with other MTZ files. This option is typically used in high resolution data collection to merge data from high medium and low resolution passes using MTZUTILS to produce a single MTZ file which is then sorted using SORTMTZ. The merged MTZ file contains a different batch number for each diffraction image, which are typically sequential for images in the same data set. The batch numbers for different data sets must be given different numbers to ensure that they are treated separately in data scaling.

Data Scaling

The program used to do all data scaling was SCALA from the CCP4 suite (CCP4 1994). An input file for SCALA was prepared which contained the batch numbers for each data set. Each of the high medium and low resolution data sets were scaled as different 'runs' using the SCALA program. This is able to put intensities recorded on different images on a single scale by allowing for variations in the intensity of the X-ray beam, as well as correcting for absorption and radiation damage. Scaling is commonly performed by determining a scale factor K and temperature factor B for each image, by refining the residual:

$$R = \sum_{h} \sum_{i} w_{hi} (I_{hi} - \langle I_h \rangle / K_{hi})^2$$

Where I_{hi} is the ith measurement of the reflection h, w_{hi} the weight for that observation (the inverse of the variance), $\langle I_h \rangle$ the weighted mean intensity for the reflection h and

$K_{hi}=K_j \exp(-2B_j \sin^2\theta_h/\lambda^2)$

K_j and B_j being the scale and B factors for the image j on which I_{hi} was measured. Furthermore θ_h is the Bragg angle for reflection h and λ the radiation wavelength.

The success of any scaling process depends on the presence of multiple symmetry related reflections on different images. It is also important when a number of non-overlapping passes have been made to ensure that there are a number of reflections in common between the passes as this will enable accurate scaling to take place. Once all the reflections have been placed on the same scale multiple observations are reduced to a single weighted mean intensity and standard deviation. At this stage it is possible to detect rogue observations or outliers providing the reflection has been measured three or more times and rejecting the offending observation(s). The proportion of rejected outliers should ideally not exceed 1 % of the data. Statistics based of the agreement between multiple observations (R_{merge}) and data completeness are calculated at this stage. The level of agreement between multiple observations can be used to modify the experimental standard deviations of the intensities estimated from Poisson statistics. Providing the multiplicity is high then the standard deviation ratio is defined as

$$SDRATIO = \left\langle \frac{I_{hi} - \langle I_h \rangle}{\sigma(I_{hi})} \right\rangle$$

Where I_{hi}, $\sigma(I_{hi})$ are the intensity and the standard deviation of the ith observation of reflection h, and $\langle I_h \rangle$ is the weighted mean intensity. SDRATIO should equal unity when averaged over a significant number of reflections. An SDRATIO greater than unity suggests that the standard deviations are underestimated. This ratio is evaluated as a function of the reflection intensity, and the standard deviations are modified by the following formula

$$\sigma'(I_{hi}) = A\sqrt{\sigma^2(I_{hi}) + BI_{hi}^2}$$

Where the values of A and B are chosen to get an SDRATIO close to unity for all intensity ranges. The resulting intensities in the scaled mtz file are then converted into amplitudes using the CCP4 program TRUNCATE. In this Bayessian statistics are implemented in order that those reflections with negative intensities can be assigned the most likely positive intensity using Wilson statistics. For refinement

of a known structure, the truncated mtz file can then be converted to SHELX (hkl) format with 5 % of the reflections being flagged for the R_{free} set, i.e. these do not go through refinement in order for the calculation of a R_{free}. Following the production of a SHELX reflection file, an instruction file (.ins file) can be generated with SHELXPRO using a pdb file containing the molecule. This ins file can be used with SHELX to refine the structure. A number of label changes and extra restraints for the non-standard residues are likely to be needed by SHELX and these must be edited into the ins file.

ESD Calculations

It is possible to determine the estimated standard deviation (ESD) associated with the bond lengths produced from high resolution unrestrained refinement. This is done by an inversion of the least squares matrix, which after scaling contains at each element i, j: c_{ij}, σ_i, σ_j with the value c_{ij} being the correlation between i and j. The standard deviations of i and j are given by σ_i, σ_j respectively. On the diagonal of the matrix i will be equal to j as the correlation of a variable with itself is one, the term thus becomes σ_i^2 and thus inverting the matrix gives the standard deviation of any parameter. In crystallography these standard deviations are termed standard uncertainties. Each atom coordinate and thermal parameter has an ESD associated with it. The radial uncertainty for an atom with co-ordinates (x, y, z) with uncertainties σ_x,σ_y,σ_z has a radial uncertainty of

$$\sigma_{radial}^2 = \sigma_x^2 + \sigma_y^2 + \sigma_z^2$$

The ESD of a bond can be calculated by noting that for two uncorrelated quantities that are subtracted the standard deviation of the resultant is

$$\sigma_{2-1}^2 = \sigma_2^2 + \sigma_2^2$$

Thus if we calculate the length of a bond we can calculate the ESD of the bond, σ_{bond} with length l from the positional uncertainties by the equation

$$\sigma_{bond}^2 = (\sigma_{x1}^2 + \sigma_{x2}^2)(\frac{x_1 - x_2}{l})^2 + (\sigma_{y1}^2 + \sigma_{y2}^2)(\frac{y_1 - y_2}{l})^2 (\sigma_{z1}^2 + \sigma_{z2}^2)(\frac{z_1 - z_2}{l})^2$$

This is the sum of the positional uncertainties projected onto the bond to account for the direction dependence of the bond. Similar considerations can be used to find the uncertainty in any quantities derived from the atomic coordinates (McRee 1999). SHELX (Sheldrick 1998) is able to calculate ESDs for proteins; first of all the structure must be refined until the R_{factor} and the R_{free} converge. Then a refinement cycle is done using the full matrix with zero shift damping and all restraints removed and using all reflection data (including the R_{free} set). A useful approximation to the full matrix calculation is the block diagonal calculation, where only portions of the full matrix are extracted into smaller matrices along the diagonal are used. A good approximation is to use a block matrix containing the x, y, z parameters for each atom without the thermal parameters. As the thermal parameters contain little information about the atomic positions, their omission however will cause underestimation of the coordinate ESD of atoms with a high B_{iso}. However the ESD values for atoms with $B_{iso} < 10$ will be correctly estimated (Cruickshank 1999). Tests done by McRee (1999a) showed that there is only a 1 % difference between calculations done with and without thermal parameters. The main advantage of blocked diagonal refinement is that it drastically reduces the computer resources needed to calculate standard uncertainties enabling the calculations to be done using 1/9 as much memory. A non-blocked diagonal refinement of endothiapepsin requires roughly 1.4 gigabytes of memory. One must also note that as the observations to parameters ratio increases (and hence the resolution) the uncertainties will decrease. At 1.6 Å resolution the average uncertainty is around 0.13 Å while at resolutions of around 1 Å the typical average uncertainty is about 0.03 Å (McRee 1999a). For bond lengths with a $B_{iso} < 10$ Å^2 and ESD values of less than 0.014 Å the diffraction data has a greater weight than the stereochemical dictionaries in refinement (Cruickshank 1999).

TLS Refinement

Three matrices can be used to describe the atomic motion of a rigid body. These are a translation matrix which describes the translation movement of the rigid body, a libration matrix which describes the rocking of the rigid group and a screw matrix which describes the screw displacement of the rigid group. They describe the anisotropic motion of rigid groups in which all the atoms in that group move as a single rigid body. These rigid bodies can be composed of a single secondary structure element, a rigid side chain such as those of the aromatic amino acids (tyrosine, phenylalanine, tryptophan and histidine) or more likely as a single protein domain. The switch from isotropic to anisotropic refinement increases the number of parameters by 2.25 fold. In contrast TLS groups can be introduced into the model for just twenty extra parameters per TLS group. This greatly helps to maintain a high ratio between observations and parameters enabling further meaningful refinement to take place. Any displacement of a rigid body can be described as a rotation about an axis that passes through a fixed point with a translation of the fixed point. When given a refined set of ADPs, TLS parameters can be generated by a least squares fit or the application of TLS parameters to a rigid body can be used to derive a single set of ADPs and hence calculated structure factors from the TLS refinement parameters. Of the three tensors T and L are symmetric while S is in general asymmetric, thus while six parameters can be used to define T and L eight parameters are needed to define S.

Winn *et al* 2001 has concluded that TLS refinement is best carried out first with the B_{iso} values fixed at a uniform value and after convergence the TLS parameters are fixed and the atomic co-ordinates and B_{iso} values are allowed to refine. It should be remembered that any hypothesis of rigidity in an atomic model can only be based on chemical sense and intuition. The Bragg reflections alone cannot support a rigid body hypothesis. The Bragg scattering does not depend on the correlation or relative phase of the motions and cannot be used to distinguish between the whole molecule rigid body motion and other possibilities. Put another way the fact that a rigid body motion fits the data does not prove there is a rigid

body motion. There will be many more non rigid body models with identical measures of fit. Refinement of TLS parameters is not at the time of writing common place and is only supported in two refinement programs RESTRAIN (Driessen *et al* 1989) and more recently REFMAC . RESTRAIN is not in widespread use and is not under active development; however REFMAC is under current development and improvement and is therefore an ideal choice for refinement using TLS groups.

TLS refinement of endothiapepsin has been conducted before on a native endothiapepsin structure using RESTRAIN (Sali *et al* 1992). The TLS matrices from this refinement were compared with 15 known structures of endothiapepsin complexed with different inhibitors. For the 15 endothiapepsin inhibitor complex structures the motions of the two domains were modelled as rigid bodies and the TLS matrices for each domain were fitted to a single screw motion along an arbitrary axis. The relative positions of the two domains could also be fitted by a screw motion (4° rotation and zero translation about an axis passing through the active site) which is consistent with the direction and magnitude of the axes determined by TLS refinement for each domain, demonstrating that the observed variation in hinge bending parameters on binding different inhibitors is consistent with the experimental TLS results. This study highlighted the flexibility of the two domains in endothiapepsin relative to each other and showed that endothiapepsin exists in two forms differing in the relative orientation of the domain comprising residues 190-302. It also highlighted the lack of interactions between the domains that makes such movements possible. One consequence of the rigid body movements highlighted in this study is the large changes in shape that occur in the S_3 pocket. This is associated with a different position and conformation of the inhibitors in the two endothiapepsin forms.

Laue Diffraction Theory

The most striking feature of a Laue diffraction pattern is the occurrence of spots forming ellipses, hyperboles and parabolas. These spot patterns are referred to as conics. All conics intersect at a single point, the position where the direct beam would have hit the detector in the absence of a beam stop. A Laue diffraction pattern consists of a large number of conics, all h, k, l planes for which hu+kv+lw=0 where u, v, w are the direct lattice indices which form a conic. A good way to show this is by the use of Ewald construction. The Ewald construction for a Laue diffraction pattern consists of three main spheres. Each of these Ewald spheres intersects at the origin (Figure 6.00).

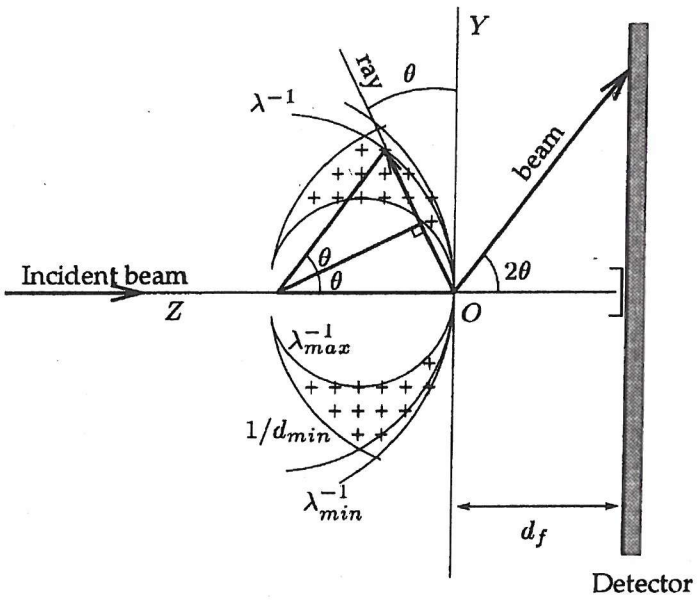


Figure 6.00 Ewald construction illustrating Laue diffraction, all RLPs between the Ewald spheres $1/\lambda_{min}$ $1/\lambda_{max}$ $1/d_{min}$ give rise to reflections (diagram taken from Ravelli 1998).

The inner sphere has a radius of $1/\lambda_{max}$ while the radius of the outer sphere is defined by $1/\lambda_{min}$. The radius of the third sphere is defined by the diffraction limit

of crystal $1/d_{min}$. For all RLPs in between these three spheres Bragg's law is fulfilled and constructive interference takes place producing reflections in the diffraction pattern. Thus the Ewald construction can be thought of as a series of spheres, one for each wavelength used in the experiment with the $1/d_{min}$ sphere defining the third edge of accessible reciprocal space. A reflecting plane passing through the observable area of reciprocal space and the origin forms a circular intersection with an Ewald sphere with a radius of $1/\lambda$. All RLP on this circular intersection give rise to diffracted rays lying on the surface of cone. The axis of this cone is defined as u, the angle between this cone axis and the incident beam is defined the semi apex ψ angle. The projection of the diffracted rays along the surface of a cone gives rise to the conics present in the diffraction pattern (Figure 6.01).

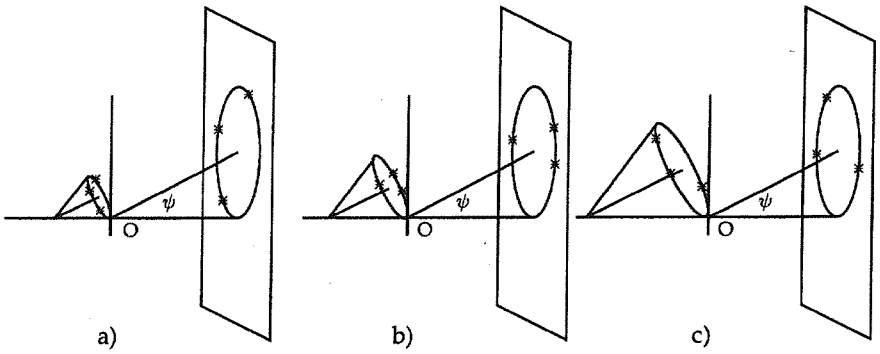


Figure 6.01 The intersection zone plane with an Ewald sphere forms a circle. The size of this circle depends on the radius of the Ewald sphere). The radius of the Ewald sphere increases from a to c, all RLPs located on these circles give rise to diffracted beams on the surface of the same cone producing a conic in the diffraction pattern (diagram taken from Ravelli 1998).

When the same reflecting plane interacts with Ewald spheres of different wavelengths the circular intersection formed is larger for shorter wavelengths. However as the cone axis angle (the semi apex ψ) remains the same, all reflections from a single reflecting plane regardless of wavelength appear on the same conical surface. The intersection between this cone and the detector forms a conic, which can be a parabola, a hyperbola, an ellipse or a straight line. A spot located where

conics overlap is called a nodal reflection. The accessible area of reciprocal space for a Laue diffraction pattern is defined as the area within the $1/\lambda_{max}$ sphere, $1/\lambda_{min}$ sphere and the $1/d_{min}$ sphere. An incoming neutron beam making an angle of θ with the plane perpendicular to the beam will enter the accessible region of reciprocal space through the internal S_I $1/\lambda_{max}$ sphere and exit through the external surface defined by the S_E $1/\lambda_{min}$ sphere and the $1/d_{min}$ sphere. The vector describing this has the name IE, the distance it travels though the accessible area of reciprocal space is dependent on θ (Figure 6.02). The maximum value that θ can have is defined as θ_m and this theta angle corresponds with an IE value of 0.

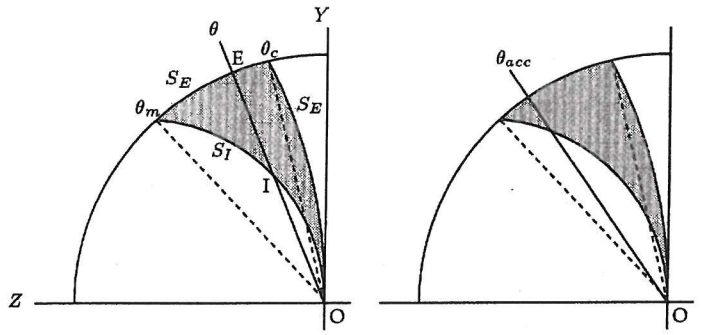


Figure 6.02 The accessible area of reciprocal space (shaded) is bounded by the internal surface S_I ($1/\lambda_{min}$) the external surface S_E ($1/\lambda_{max}$) and $1/d_{min}$). The length of the vector IE that crosses S_I and S_E is defined by θ. θ_c is minimum possible θ angle and θ_m is maximum θ angle with θ_{acc} being the maximum θ angle that the detector can record (diagram taken from Ravelli 1998).

Spatial overlap

The spots on a Laue diffraction pattern have a certain size that is dependent mainly on the mosaic spread of the crystal. When a spot encroaches on its nearest neighbour it is classed as a spatial overlap. The minimum value for the θ angle is called θ_c at this angle the length of IE is at its maximum. Thus the density of reflections in the diffraction pattern is highest when $\theta=\theta_c$. Thus when $\theta=\theta_c$ there is a greater probability of spatial overlap of reflections. Also most overlaps occur between singlets making it critical to try and reduce spatial overlap to a minimum.

In most experiments the detector is not able to record all θ angles, the maximum θ value that can be sampled in the observable area of reciprocal space is called θ_{acc}. For nearly all detectors θ_{acc} lies close to θ_m. Increasing the distance between the detector and crystal can be used to reduce the spatial overlap of reflections.

Harmonic overlaps

There is second condition producing overlapping reflections in Laue diffraction, harmonic overlap is often called energy overlap and unlike spatial overlap it is completely independent of the crystal to detector distance. Harmonic overlaps can be composed of several reflections with each reflection corresponding to a different wavelength e.g.

d	λ
d/2	λ/2
d/3	λ/3

This harmonic overlapping of reflections can be seen quite clearly from Bragg's law, which states that.

$2d\sin\theta=\lambda n$

In a monochromatic experiment the λn side of the equation is constant as a single wavelength is used. However in a Laue experiment for a family of lattice planes with spacing d, Bragg's law is simultaneously satisfied by any wavelength which is a multiple of the spacing between the lattice planes. Hence a series of lattice planes (with spacing d) which in a set orientation produce a reflection with X-rays or neutrons with a wavelength of 1.3 Å (n1) will also produce a reflection with X-rays or neutrons with a wavelength of 2.6 Å (n2). Thus Bragg's law is simultaneously satisfied by all wavelengths for which

$(n\lambda/n)$

All orders of n for the spacing d directly superimpose increasing the intensity of the given reflection. While any RLPs in the $1/d_{min}$ resolution sphere separated by $(n\lambda/n)$ would superimpose, only a small number of these are actually in the observable area of reciprocal space. An exhaustive analysis of harmonic overlaps is given by Cruickshank *et al* 1987. One of the major conclusions reached is that the highest probability of observing single reflections occurs at longer wavelengths and higher resolutions (larger θ values). Also no singles can be observed for RLPs corresponding to wavelengths longer than $2\lambda_{min}$ and resolutions lower than $2d_{min}$. These observations mean that a large proportion of the low-resolution reflections are present in harmonic overlaps and not as single spots. It also indicates that the θ cut off value defined by θ_{acc} is very important, the closer θ_{acc} is to θ_m the better as this area of reciprocal space is the most likely to give single reflections. The Laue geometry is very sensitive to the mosaic spread of the crystal; a mosaic spread larger than 0.3 degrees causes radial elongation of reflections. This can be a major problem in the Laue study of micro crystals, however this was not a major problem in this series of experiments.

Nodal Reflections

One striking feature of a Laue diffraction pattern are the spots at which different conics intersect, these spots are the nodal reflections. They are spots with low hkl values or multiples thereof and are associated with principal zones of the lattice such as hk0 (Helliwell 1992). Thus nodals are formed by the harmonic overlap of reflections however the nearest neighbour to a nodal spot is always a single. These spots have a h' k' l' index related to the highest integer indices in the nodal reflection such as h'=(n+1)h so that h' k' l' cannot have a common integer divisor and h' k' l' must be stimulated by a single wavelength. A survey of several nearest neighbouring spots to nodals in a Laue diffraction pattern can therefore be used to determine the resolution limit of the data from the crystal (Helliwell 1992).

References

Andrevta, N.S., Rumsh L.D. (2001) Analysis of crystal structures of aspartic proteinases: On the role of amino acid residues adjacent to the catalytic site of pepsin-like enzymes. Protein Science 10 2439-2450

Arndt, U. W. in Rossman, M.G., Arnold E. (2001) International tables for crystallography Volume F: crystallography of biological macromolecules. 125-132 Kluwer academic publishers.

Bailey, D., Cooper, J. B., Veerapandian, B., Blundell, T. L., Atrash, B., Jones, D. M. & Szelke, M. (1993) X-ray crystallographic studies of complexes of pepstatin A and a statine-containing human renin inhibitor with endothiapepsin. Biochem. J. 289, 363-371.

Barkholt, V. (1987) Amino acid sequence of endothiapepsin. Complete primary structure of the aspartic protease from *Endothia parasitica*. Eur. J. Biochem. 167: 327-38.

Blundell, T. L., Cooper, J. B., Foundling, S. I., Jones, D. M., Atrash, B. & Szelke, M. (1987) On the rational design of renin inhibitors: X-ray studies of aspartic proteinases complexed with transition state analogues. Biochemistry 26, 5585-5590.

Blundell, T.L., Jenkins, J. A., Sewell, B. T., Pearl, L. H., Cooper, J. B., Tickle, I. J., Veerapandian, B. & Wood, S. P. (1990) X-ray analyses of aspartic proteinases. The three-dimensional structure at 2.1 A resolution of endothiapepsin. J Mol Biol 211:919-941.

Beveridge, A.J. (1998) A theoretical study of the initial stages of catalysis in the aspartic proteinases. J. Mol. Struct. THEOCHEM, Volume 453, 275-291

Bon, C., Lehmann, M.S. and Wilkinson, C. (1999). Quasi-Laue neutron diffraction study of the water arrangement in crystals of triclinic lysozyme from hen egg-white. Acta Crystallogr. *D55*, 978-987.

Brünger A.T., (1992) The Free R Value: a Novel Statistical Quantity for Assessing the Accuracy of Crystal Structures, Nature 355, 472-474

Burmeister, W. P. (2000) Structural changes in a cryo-cooled protein crystal owing to radiation damage. Acta Crystallogr. D56, 328-341.

Campbell, J. W., Hao, Q., Harding, M. M., Nguti, N. D. & Wilkinson, C. (1998). LAUEGEN version 6.0 and INTLDM. J. Appl. Cryst. 31, 496-502.

Carrell, H. L., and Glusker J. P. in Rossman, M.G., Arnold E. (2001) International tables for crystallography Volume F: crystallography of biological macromolecules. 111-116 Kluwer academic publishers.

Cassidy, C.S., Lin J., Frey P.A. (1997) A new concept for the mechanism of action of chymotrypsin; The role of the low barrier hydrogen bond. Biochemistry 36, 4576-4584

Choi, G. H., Pawlyk, D. M., Rae, B., Shapira, R. & Nuss, D. L. (1993) Molecular analysis and overexpression of the gene encoding endothiapepsin, an aspartic protease from *Cryphonectria parasitica*. Gene 125: 135-41 (1993)

Cleland, W. W., Frey, P. A. and Gerlt, J. A. (1998). The low barrier hydrogen in enzymatic catalysis. J. Biol. Chem. 273, 25529-25532.

Coates L., Erskine P.T, Crump M.P, Wood S.P. and Cooper J.B. (2002) Five atomic resolution structures of Endothiapepsin inhibitor complexes: Implications for the aspartic proteinase mechanism. J. Mol. Biol. 318 1405-1415

Coates L., Erskine P.T, Wood S.P., Myles D. and Cooper J.B. (2001) A Neutron Laue diffraction study of endothiapepsin; implications for the aspartic proteinase mechanism. Biochemistry 40, 13149-13157

Collaborative Computational Project Number 4 (1994). "The CCP4 suite: programs for protein crystallography", Acta Cryst. D50, 760-763.

Cooper J.B. (2002) Aspartic proteinases in disease: a structural perspective. Current drug targets 3, 155-174

Cooper, J. B., Foundling, S. I., Blundell, T. L., Boger, J., Jupp, R. A., & Kay, J. (1989) X-ray studies of aspartic proteinase-statine inhibitor complexes. Biochemistry 28, 8596-8603.

Cooper, J. B., Foundling, S., Hemmings, A., Blundell, T., Jones, D. M., Hallett, A. & Szelke, M. (1987) The structure of a synthetic pepsin inhibitor complexed with endothiapepsin. Eur. J. Biochem. 169, 215-221.

Cooper, J. B., Quail, W., Frazao, C., Foundling, S. I., Blundell, T. L., Humblet, C., Lunney, E. A., Lowther, W. T. & Dunn, B. M. (1992) X-ray crystallographic analysis of inhibition of endothiapepsin by cyclohexyl renin inhibitors. Biochemistry 31, 8142-8150.

Cooper, J.B., & Myles, D. A. A. (2000) A preliminary Neutron Laue diffraction study of the aspartic proteinase endothiapepsin. Acta. Cryst. D56 246-248

Cronin, N.B., Badasso, M.O., Tickle, I.J., Dreyer, T., Hoover, D.J., Rosati, R.L., Humblet, C.C., Lunney, E.A. & Cooper, J.B. (2000) X-ray structures of five rennin inhibitors bound to Saccharopepsin: Exploration of active site specificity. J. Mol. Biol. 303 745-760

Cruickshank, D. W. J.,Helliwell, J. R., and Moffat, K. (1987) Multiplicity distrubution of reflections in Laue diffraction. Acta Cryst., A47, 352-373

Cruickshank, D. W. J. (1999). Remarks about protein structure precision. Acta Cryst. D55, 583-601.

Deacon A., Gleichmann T., Kalb A.J., (Gilboa), Price H., Raftery J., Bradbrook G., Yariv J. and Helliwell J.R. "The structure of concanavalin A and its bound solvent determined with small-molecule accuracy at 0.94Å resolution" (1997) Faraday Transactions, 93 (24), 4305-4312.

Dealwis C., (1993) Crystallographic analysis of the binding of a phosphostatine renin inhibitor by an aspartic proteinase. PhD Thesis, University of London

Drenth, J. (1994) Principles of protein X-ray crystallography, Springer-Verlag.

Driessen, H., Haneef, M. I. J., Harris, G. W., Howlin, B., Khan, G. & Moss, D. S. (1989). RESTRAIN: restrained structure-factor least-squares refinement program for macromolecular structures. J. Appl. Cryst. 22, 510-516.

Dunn, B. M. & Kay, J. (1985) Design, synthesis and analysis of synthetic substrates for aspartic proteinases. Biochem Soc Trans 13, 1041-1043.

Dunn, B.M. (1991) Structure and function of the aspartic proteinases. Plenum press, New York.

Leslie, A. G. W. in Fanchon, E., Geissler, E., Hodeau.J., Regnard, J., Timmins, P.T., (2000) Structure and dynamics of biomolecules. 14-35 Oxford university press

Fassbender, K., Masters, C. and Beyreuther, K. (2001). Alzheimer's disease: molecular concepts and therapeutic targets. Naturwissenschaften, 88, 261-267.

Foundling, S. I., Cooper, J. B., Watson, F. E., Pearl, L. H., Sibanda, B. L., Wood, S. P., Blundell, T. L., Valler, M. J.,Norey, C. G., Kay, J., Boger, J., Dunn, B. M., Leckie, B. J., Jones, D. M., Atrash, B., Hallett, A. & Szelke, M. (1987) High resolution X-ray analysis of renin inhibitor aspartic proteinase complexes. Nature 327, 349-352.

Frey, P. A., Whitt, S., and Tobin, J. (1994) A low-barrier hydrogen bond in the catalytic triad of serine proteases. *Science* 264, 1927-1930
Fusek, M. and Vetvicka, V. (1994). Mitogenic function of human procathepsin-D - the role of the propeptide. Biochem. J., 303, 775-780.

Fruton J.S. (1976) The mechanism of the catalytic action of pepsin and related acid proteinases. Adv. Enzymol. Relat. Areas Mol. Biol. 44 1-36

Garman, E. (1999). Cool data: quantity and quality. Acta Cryst. D55, 1641-1653

Gruner, S. M., Eikenberry E. F., and Tate M. W. in Rossman, M.G., Arnold E. (2001) International tables for crystallography Volume F: crystallography of biological macromolecules. 143-154 Kluwer academic publishers.

Habash, J., Raftery, J., Nuttall, R., Price, H. J., Wilkinson, C., Kalb (Gilboa), A. J. and Helliwell, J. R. (2000). Direct determination of the positions of the deuterium atoms of the bound water in concanavalin A by neutron Laue crystallography. *Acta Crystallogr. D56*, 541-550.

Ha, N., Choi, G., Yong Choi K. and Oh, B., Structure and enzymology of Δ5-3-ketosteroid isomerase, (2001) Current Opinion in Structural Biology, 11, 674-678.

Helliwell, J.R. (1988) Protein Crystal Perfection and the nature of radiation-damage Journal of Crystal Growth 90 259-272

Helliwell J.R., (1992) Macromolecular crystallography with synchrotron radiation. Cambridge University Press.

Hong, L., Koelsch, G., Lin, X. L., Wu, S. L., Terzyan, S., Ghosh, A. K., Zhang, X. C., Tang, J. (2000). Structure of the protease domain of memapsin 2 (β-secretase) complexed with inhibitor. Science, 290, 150-153.

James, M. N. G., Sielecki, A. R. (1983) Structure and refinement of pencillopepsin at 1.8 Å resolution J. Mol. Biol. 163 299-361

James, M. N. G., Sielecki, A. R., Hayakawa, K., Gelb, M. H. (1992) Crystallographic analysis of transition state mimics bound to penicillopepsin: difluorostatine- and diflurostatone-containing peptides. Biochemistry 31, 3872-3886.

Jenkins, J.A., Blundell, T. L., Tickle, I. J. & Ungaretti, L. (1975) The low resolution structure analysis of an acid proteinase from *Endothia parasitica*. J. Mol. Biol. 99, 583-590.

Kabsch,W. and Sander,C. (1983) Biopolymers 22, 2577-2637

Kabsch,W. (1993), Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants J. Appl. Cryst. 26, 795-800.

Kossiakoff, A. A. and Spencer, S. A. (1980). Neutron diffraction identifies His 57 as the catalytic base in trypsin. Nature 288, 414-416.

Kossiakoff, A. A. and Spencer, S. A. (1981). Direct determination of protonation states of aspartic acid-1-2 and histidine-57 in the tetrahederal intermediate of the serine proteases: neutron structure of tryspin. Biochemistry 20, 6462-6474.

Leslie, A.G.W., (1992), Joint CCP4 + ESF-EAMCB Newsletter on Protein Crystallography, No. 26.

Ladd , M.F.C., Palmer R.A., (1993) Structure determination by X-ray Crystallography (Third edition) Plenum Press.

Lin, Y., Fusek, M., Lin, X., Hartsuck, J.A., Kezdy, F.J. and Tang. J. (1992) J. Biol. Chem. 267 18418

Lunney, E. A., Hamilton, H. W., Hodges, J. C., Kaltenbrohn, J. S., Repine, J. T., Badasso, M., Cooper, J. B., DeAlwis, C., Wallace, B., Blundell, T. L., Lowther, W. T., Dunn, B. M. & Humblet, C. (1993) The analysis of five endothiapepsin crystal complexes and their use in the design and evaluation of novel renin inhibitors. J. Med. Chem. 36, 3809-3820.

McDermott, A., Ridenour, C.F., (1996) Encyclopedia of NMR. Wiley

McRee, D.E. (1999) XtalView/Xfit - A Versatile Program for Manipulating Atomic Coordinates and Electron Density. Journal Structural Biology, vol. 125, pp. 156-165.

McRee D., Practical protein crystallography (1999a) (Second Edition). Academic press

Merrit E.A., Bacon D.J. (1997) Raster 3D: Photorealistic molecular graphics. Methods in Enzymology 277 505-524

Merritt E.A., (1999) Expanding the model: anisotropic displacement parameters in protein structure refinement. Acta. Cryst. D D55, 1109-1117

Merritt E.A., (1999a) Comparing anisotropic displacement parameters in protein structures. Acta. Cryst. D D55, 1997-2004

Moews, P. and Bunn, C. W. (1970) An X-ray crystallographic study of the rennin-like enzyme of Endothia parasitica. J. Mol. Biol. 54, 395-397.

Myles, D., Bon, C., Langan, P., Cipriani, F., Castagna, J., Lehmann, M., Wilkinson, C. (1998) Neutron Laue diffraction in macromolecular crystallography Physica B 241 1122-1130

Nicholls, A., Sharp, K. and Honig, B. (1991) PROTEINS, Structure, Function and Genetics. 11 281-291

Niimura, N., Minezaki, Y., Nonaka, T., Castagna, J., Cipriani, P., Lehmann M & Wilkinson, C. (1997) Neutron Laue diffractometry with an imaging plate provides an effective data collection regime for neutron protein crystallography. Nature Structural Biology 4, 11, 909-914

Ondetti, M. A. and Cushman, D. W. (1982). Enzymes of the renin-angiotensin system and their inhibitors. Annu. Rev. Biochem., 51, 283-308.

Pearl, L. & Blundell, T. (1984) The active site of aspartic proteinases. FEBS Lett. 174, 96-101.

Piot, P., Bartos, M., Ghys, P. D., Walker, N. and Schwartlander, B. The global impact of HIV/AIDS. Nature (Lond.), 410, 968-973.

Ravelli, R., (1998) Laue diffraction and fast monochromatic X-rat data collection techniques. PhD Thesis, University of Utrecht.

Ravelli, R. B. G. and McSweeney, S. M. (2000). The fingerprint that X-rays leave on structures. Structure 8, 315-328.

Razanamparany, V., Jara, P., Legoux, R., Delmas, P., Msayeh, F., Kaghad, M., & Loison, G. (1992) Cloning and mutation of the gene encoding endothiapepsin from *Cryphonectria parasitica*. Curr Genet 21: 455-61.

Rhodes G., Crystallography made crystal clear (Second Edition). Academic press 2000

Rodrigues, E. J., Angeles, T. S. and Meek, T. D. (1993). Use of nitrogen-15 kinetic isotope effects to elucidate details of the chemical mechanism of human immunodeficiency virus protease. Biochemistry *32*, 12380-12385.

Sali, A., Veerapandian, B., Cooper, J. B., Foundling, S. I., Hoover, D. J. & Blundell, T. L. (1989) High resolution X-ray study of the complex between endothiapepsin and an oligopeptide inhibitor: The analysis of inhibitor binding and description of the rigid body shifts in the enzyme. EMBO J. 8, 2179-2188.

Sali, A., Veerapandian, B., Cooper, J. B., Moss, D. S., Hofmann, T. and Blundell, T. L. (1992) Domain flexibility in aspartic proteinases. Proteins Struct. Func. Genet. 12, 158-170.

Schoenborn, B. P., and Knott R. in Rossman, M.G., Arnold E. (2001) International tables for crystallography Volume F: crystallography of biological macromolecules. 133-142 Kluwer academic publishers.

Serpell, L.C., Blake, C.C.F. and Fraser, P. E. (2000). Molecular structure of a fibrillar Alzheimer's Aβ fragment. *Biochem.*, 39, 13269-13275.

Shan, S., Loh, S. Herschlag, D. The energetics of Hydrogen bonds in model systems: Implications for enzymatic catalysis (1996) Science 272, 97-101

Sheldrick,G.M. in direct methods for solving macromolecular structures, (1998) pp.131-141 and 401-411 Oxford university press, Oxford.

Sherwood, D. Crystals, X-rays and proteins (1976) Longman Group Ltd, London

Steiner, T., and Saenger, W. (1994) *Acta Crystallogr. Sect. B Struct. Sci.* **50**, 348-357

Stephenson RC, Clarke S. 1989. Succinimide formation from aspartyl and asparaginyl peptides as a model for the spontaneous degradation of proteins. *J. Biol. Chem.* **264**, 6164-6170.

Stratton, J.R., Pelton, J.G., and Kirsch J.F., (2001) A novel engineered subtilisin BPN' lacking a low-barrier hydrogen bond in the catalytic triad. Biochemistry 40 10411-10416.

Sturmann, H. B. in Fanchon, E., Geissler, E., Hodeau.J., Regnard, J., Timmins, P.T., (2000) Structure and dynamics of biomolecules. 390-406 Oxford university press

Subramanian, E., Swan, I. D. A., Liu, M., Davies, D. R., Jenkins, J. A., Tickle, I. J. & Blundell, T. L. (1977) Homology among acid proteases: comparison of crystal structures at 3 Angstrom resolution Proc. Natl. Acad. Sci. USA 74, 556-559.

Suguna, K., Padlan, E. A., Smith, C. W., Carlson, W. D. & Davies, D. (1992) Binding of a reduced peptide inhibitor to the aspartic proteinase from Rhizopus chinensis: implications for a mechanism of action. Proc. Natl. Acad. Sci. USA 84, 7009-7013.

Tomasselli, AG & Heinrikson, RL (2000). Targeting the HIV-protease in AIDS therapy: a current clinical perspective. *Biochem. Biophys. Acta*, **1477**, 189-214.

Umezawa, H., Aoyagi, T., Morishima, M., Matsuzaki, M., Hamada, M. and Takeuchi, T. (1970) Antibiotics 23, 259-261.

Veerapandian, B., Cooper, J., Sali, A. & Blundell, T. L. (1990) Three dimensional structure of endothiapepsin complexed with a transition-state isostere inhibitor of renin at 1.6 Angstroms resolution. J. Mol. Biol. 216, 1017-1029.

Veerapandian, B., Cooper, J. B., Sali, A., Blundell, T. L., Rosatti, R. L., Dominy, B. W., Damon, D. B. & Hoover, D. J. (1992) Direct observation by X-ray analysis of the tetrahedral "intermediate" of aspartic proteinases. Protein Sci. 1, 322-328.

Vetvicka, V., Vetvickova, J. and Fusek, M. (1999). Anti-human procathepsin D activation peptide antibodies inhibit breast cancer development. Breast Cancer Res. Treat., 57, 261-269.

Wang, Z., Luecke, H., Yao, N. & Quiocho, F.A. (1997) A low energy short hydrogen bond in very high resolution structures of protein receptor-phosphate complexes. Science 4 519-522

Weik, M., Ravelli R. B. G., Kryger G., McSweeney S., Maria L. Raves M.L.,, Michal Harel M., Gros P., Silman I., Kroon J., and Sussman J.L. (2000) Specific chemical and structural damage to proteins produced by synchrotron radiation. PNAS 97 623-628

Williams, D. C., Whitaker J. R. & Caldwell, P. V. (1972) Hydrolysis of peptide bonds of the oxidised B-chain of insulin by Endothia parasitica protease. Arch. Biochem. Biophys. 149, 52-61.

Winn, M. D., Isupov, M. N. & Murshudov, G. N. (2001). Use of TLS parameters to model anisotropic displacements in macromolecular refinement. Acta Cryst. D57, 122-133.