UNIVERSITY OF SOUTHAMPTON

Crystallographic and biochemical analysis of three distinct hydrolases: *Dermatophagoides pteronyssinus* 1 (*Der p*1), momordin and the bacterial carbon-carbon hydrolase, MhpC

By Graham Spencer Dunn

A thesis presented for the degree of

DOCTOR OF PHILOSOPHY

Department of Biochemistry

September 2000

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SCIENCE

BIOCHEMISTRY

Doctor of Philosophy

CRYSTALLOGRAPHIC AND BIOCHEMICAL ANALYSIS OF THREE DISTINCT HYDROLASES: DERMATOPHAGOIDES PTERONYSSINUS 1 (DER P1), MOMORDIN AND THE BACTERIAL CARBON-CARBON HYDROLASE, MHPC

by Graham Spencer Dunn

MhpC is one of six enzymes derived from the mhp operon, which collectively give rise to the meta-cleavage pathway for the degradation of 3-(3-hydroxyphenyl)propionic acids (3-HPP) in Escherichia coli. These enzymes work in tandem to convert 3-HPP to metabolites which can then be fed into the Krebs' cycle for energy release. The role of MhpC in this pathway is to convert 2-hydroxy-6-ketonona-2,4-diene-1,9-dioate to 2-hydroxy-penta-2,4-dienoate and succinate. This reaction involves the breaking of a carbon-carbon bond, which constitutes a rare catalytic event in nature. Elucidating the catalytic mechanism of MhpC will provide invaluable information at the molecular level contributing to the understanding of how biodegradation of environmental pollutants can be achieved.

The MhpC structure has been solved using Multiwavelength Anomalous Dispersion (M.A.D.) phasing to obtain the protein phase information from 32 selenomethionine sites in the MhpC asymmetric unit. The enzyme belongs to the α/β hydrolase fold family. Mechanistic studies can now be undertaken involving the analysis of the interactions of various ligands bound in at the active site.

Momordin is a ribosome inactivating protein (RIP) from the seeds of the bitter gourd, Momordica charantia. RIPs exist in two main classes. Type II RIPs, such as ricin, possess two subunits, the A- and B-chains. The A-chain is a toxin which targets ribosomes, whilst the B-chain is a lectin responsible for facilitating A-chain entry into cells. Type I RIPs, in contrast, are monomeric toxins, which share homology with type II RIP A-chains. The key to the toxicity of these proteins lies in their highly specific N-glycosidase activity. This involves the cleavage of a ribosomal RNA (rRNA) adenine-ribose glycosidic bond found in a critically-conserved adenosine (position A4324 in rat 28S rRNA) present in a region of the rRNA termed the sarcin/ricin loop, which is essential for elongation factor binding to the ribosome. This single cleavage event is responsible for complete inactivation of the ribosome.

The purpose of the crystallographic study on momordin presented here was to provide mechanistic detail about the functioning of momordin through ligand binding studies.

Momordin crystals were soaked in βNADH and data was then collected to 2Å resolution. The resultant electron density maps showed that the ligand's nicotinamide group had been cleaved in preference to the adenine. The reasons for this are unclear, and further ligands will be analysed in the future to provide more information about this unusual behaviour.

Der p1 is a proteolytic enzyme derived from the house dust mite, Dermatophagoides pteronyssinus. It has been identified as a major allergen involved in asthma linked to house dust. Once inhaled by those susceptible to house dust allergy, Der p1 then passes into the bloodstream, where it initiates an immune system cascade. The direct result of this is the release of cytokines, including histamine, which bring about airway inflammation, and asthma. The specific role which Der p1 performs in asthma is complex and thus is not fully understood, but it is likely that both its proteolytic activity and surface epitopes are involved.

The aim of the work with Der p1 was to express, purify and crystallise the allergen, and to then solve its crystal structure. From this it was hoped, inhibitors could be designed for household use, which block the protease's digestive activities. The allergen's structure could also serve as the basis for designing mutants for immunotherapy.

High level expression of proDer p1-thioredoxin fusion protein was achieved. The recombinant enzyme was stored in an insoluble form in the cell. Refolding of proDer p1-thioredoxin fusion protein was successful at low levels.

CONTENTS:

ACKNOWLEDGEMENTS

I would like to begin by thanking my supervisors, Prof. Steve Wood and Dr. Jon Cooper, for all of their invaluable guidance and help throughout the past four years!

My gratitude is further extended to the BBSRC and Wessex Medical Trust for their funding of this work.

Thanks must also go to Prof. Noor Kalsheker and Louise McDonald for their support with the *Der p*1 project, and also to Prof. Tim Bugg and Tom Robertson, for their help with MhpC.

I would like to thank and acknowledge others in the department, who over the years have offered their advice and assistance with the work presented in this thesis. In this regard, I would like to mention the following people: Darren, Iain, John Reid, Maru, Raj and Shu-fen. In addition, I would like to give a special acknowledgement to Alun for being incredibly supportive and helpful throughout the entire PhD!

It is important to mention some of my other colleagues in the workplace who have made these years so memorable. So thanks to: Alan, Bob, Doug, Ed, Gareth, Helen Jameson, Helen Rodway, Jo Robinson, Julie, Michelle, Neil Johnson, Neil Jones, Niamh, Nicola Preston, Noom, Rob, Robert, Richard Duggleby, Richard Fowler, Tiffany and Tracey.

Finally, I would like to add a deeply personal thank you to **all** of my family and friends who have supported me over the years, thinking especially of my mum, dad, sister, grandparents and, of course, Fiona. Thank you for all of your love and encouragement along the way!

This thesis is dedicated to those who created, and educated, me.

ABBREVIATIONS

3-HPP	3-(3-hydroxyphenyl)propionic acid
Å	Ångström (10^{-10}m)
A	Adenine
ABR	Abrin A chain
ADP	Adenosine diphosphate
AIDS	Acquired immunodeficiency syndrome
Ala	Alanine
Arg	Arginine
Asp	Aspartic acid
ATP	Adenosine triphosphate
BaFBr	Barium fluorobromide
BphD	2-hydroxy-6-oxo-6-phenylhexa-2,4-dienoate hydrolase
BPI	Barley protein synthesis inhibitor
^{o}C	Celsius
C	Cytosine
CCD	Charge-coupled-device
CCP4	Collaborative Computing Project number 4
CD4	Cluster determination factor 4
Cys	Cysteine
Da	Dalton
***Der p*1**	*Dermatophagoides pteronyssinus* 1
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
DTT	Dithiothreitol
E. Coli	*Escherichia coli*
EDTA	Ethylenediaminetetraacetic acid
ER	Endoplasmic Reticulum
ESRF	European Synchrotron Radiation Facility
$Eu^{2+/3+}$	Europium ion
FMP	Formycin 5'-monophosphate

G	Guanine
GeV	Giga electron volts
Glu	Glutamic acid
Gly	Glycine
GM-CSF	Granulocyte/macrophage- colony stimulating factor
Gu-HCl	Guanidinium hydrochloride
HEPES	N-(2-hydroxyethyl)piperazine-N'-(2-ethanesulphonic) acid
His	Histidine
HIV	Human immunodeficiency virus
IgE	Immunoglobulin E
IL-8	Interleukin-8
Ile	Isoleucine
IPTG	Isopropyl-β-D-thiogalactopyranoside
k_{cat}	Catalytic constant
kDa	Kilodalton
K_M	Michaelis constant
kV	Kilovolt
kW	Kilowatt
LB	Millers Luria Broth
Leu	Leucine
LINAC	Linear accelerator
LUF	Luffin-α
μm	Micrometre
μM	Micromolar
M	Molar
mA	Milliampere
MAD	Multiwavelength Anomalous Diffraction
MES	2-(N-morpholino)ethanesulphonic acid
Met	Methionine
Mg/ml	Milligrams per millilitre
Mhp	Refers to the *meta*-cleavage pathway for 3-(3-**h**ydroxyphenyl) **p**ropionic acid in *E. coli*

MhpC	2-hydroxy-6-ketonona-2,4-diene-1,9-dioate-5,6-hydrolase
MIR	Multiple isomorphous replacement
MMC	α-momorcharin (Momordin I)
MPD	2-methyl-2,4-pentanediol
mRNA	Messenger ribonucleic acid
NAD^+	Nicotinamide adenine dinucleotide (oxidised form)
NADH	Nicotinamide adenine dinucleotide (reduced form)
NADP	Nicotinamide adenine dinucleotide phosphate
nm	Nanometre
NMM	New minimal medium
NMR	Nuclear Magnetic Resonance
Nu	Nucleophile
OD	Optical density
OD_{600}	Optical density at 600 nanometres
PAP	Pokeweed antiviral protein
PEG	Polyethylene glycol
Phe	Phenylalanine
pI	Isoelectric point
PMSF	Phenylmethylsulphonylfluoride
***Pro Der p*1**	*Pro Dermatophagoides pteronyssinus* 1
RIP	Ribosome inactivating protein
RMS	Root mean squared
RNA	Ribonucleic acid
RNase	Ribonuclease
RPM	Revolutions per minute
rRNA	Ribosomal ribonucleic acid
RTA	Ricin A chain
SAP	Saporin SO-6
SDS	Sodium dodecyl sulphate
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
Ser	Serine
Sm	Small amino acid

SP	Sulphopropyl
SP-A	Surfactant protein A
SP-D	Surfactant protein D
SRL	Sarcin/ricin loop
T	Thymine
TCS	Trichosanthin
Thr	Threonine
TNFα	Tumour necrosis factor alpha
tRNA	Transfer ribonucleic acid
Tyr	Tyrosine
UV	Ultraviolet
V	Volts
Val	Valine
X	Any amino acid

Chapter 1

Introduction

1. Introduction:

1.1 Asthma and *Dermatophagoides pteronyssinus* 1

In many so-called Westernised societies we have, for many years now, seen an increase in the prevalence of asthma (a respiratory disorder characterised by difficulty in breathing). The onset of an asthma attack typically begins with the inhalation of triggering molecules (allergens) into the respiratory tract. These allergens then pass into the bloodstream and, when in sufficient concentration, bring about an elevation in blood IgE levels, which in turn leads to an immune system cascade and the production of large quantities of numerous cytokines such as histamine which induce airway inflammation leading to wheezing.

Certain individuals are known to be genetically predisposed to asthma and for these the likelihood of developing the disorder is thought to decrease with the ageing process. The initial onset in these individuals is thought to be linked with exposure to so-called inducers such as indoor allergens [1]. Once sensitised to these allergens, susceptible individuals are then sensitive to both "triggers" (e.g. exercise and cold air), which are responsible for bringing about the inflammation and subsequent airway restriction associated with asthma, and "enhancers" (e.g. rhinovirus and ozone) which can lead to an increase in the severity of the disorder [1].

In recent decades, one of the most striking developments made to our lives has been the shift towards a more comfortable indoor environment, and it was this observation which led scientists to postulate that indoor allergens played a major role in asthma since the incidence of the disorder increased equally rapidly during the same time period [1]. Since this suggestion was made numerous indoor allergens have been identified including those shown below (see Table 1):

Organism	Example of Allergen	Further Information
House Dust Mite	*Der p*1 *Der f*2	-Key player in asthma (see below). -Evidence for a role in dermatitis [2,3] -NMR structure solved: single domain protein of immunoglobulin fold [4]
Cat/Dog (Animal Dander)	*Fel d*1/ *Can f*1	-Produced by sebaceous glands and shed from fur
Cockroach	*Bla g*2	-Shares homology with aspartic proteases -Digestive enzyme excreted by cockroach

Table 1. Indoor allergens and their sources [1]

Some of the most important sources of indoor allergens are the house dust mites, and it is to these organisms that we shall now turn our attentions. The aforementioned recent move toward an indoor environment has greatly favoured mite survival and proliferation. There are three primary determinants of mite growth: humidity (>60%), temperature (>21°C) and food supply (human skin scales) [1]. Now, modern housing has improved insulation (e.g. double glazing) and has provided upholstered furniture and carpets, and these help to maintain an internal environment of high humidity and/or of reasonably warm temperature [1].

So, the shift towards a comfortable indoor environment has favoured mite growth, but what is the significance of this? The house dust mite, *Dermatophagoides pteronyssinus*, produces several allergens, the most abundant of which is *Dermatophagoides pteronyssinus* 1 (*Der p*1) which has been shown to react with Immunoglobulin E (IgE) in the sera of 80% of patients suffering from house dust allergy [5,6].

Attempts to control the allergens and their effects have yielded varying degrees of success, allergen avoidance can be effective, which can be achieved by covering bedding with allergen-impermeable materials, whilst treatment of carpets and furniture with chemicals such as denaturants and acaricides is only moderately effective due to their lack of penetrating

power and the persistence of viable allergens. Vacuuming carpets can give limited benefit since this has little effect on live mites and sprays faecal particles through the collection bag and out into the room air. However, vacuum cleaners with double-thickness collection bags or filters are significantly better. An alternative approach to countering the allergens' effects is via clinical methods such as immunotherapy. These methods are currently a main focus of attention for researchers dealing with asthma. Repeated injections of allergen extracts have proven effective for seasonal asthma, but have been less successful for perennial asthma. Alternative versions of this immunotherapy approach include the usage of small allergen peptides which could specifically react with T-cells but not with IgE, and the usage of genetically-modified allergens which, despite possessing over 90% of the wild-type primary structure, no longer bind IgE [1].

Let us now briefly scan over the history of the *Der p*1 allergen since it is quite intriguing: In 1698, J. Floyer uncovered the earliest reported link between asthma and exposure to house dust [7]. This was a pioneering observation which was followed up more than two centuries later by P.A. Kern in 1921 [8] and Spivacke, *et al*, in 1925 [9], who produced the first reports of positive immediate skin prick tests with house dust extracts and asthmatics. Then, in 1964, Voorhorst, *et al*, suggested that the house dust allergens arose from house dust mites [10], and this hypothesis was later supported in 1967 by the same team, who showed a correlation between house dust mite numbers and allergenic potency of dust extracts [11].

The next breakthrough came in 1980 when Chapman, *et al*, discovered a major house dust mite allergen which they succeeded in purifying and characterising and which they ultimately named *Der p*1 [12]. This, in turn, led to the discovery in 1981 by Tovey, *et al*, that the allergen was primarily present at high concentration in mite faecal balls, and thus was an excreted protein [13]. Then, in 1988, Chua, *et al*, isolated a cDNA clone encoding *Der p*1 which they then sequenced [14]. Upon inspection of the inferred amino acid sequence derived

from the cDNA, it was reported that the mature 222 residue *Der p*1 shared 21-29% homology with several cysteine proteases, viz. papain, actinidin, and cathepsin-B and -H [14] (although it has now been shown that there was a small error in the original sequence and as such these homology values are now slightly altered (see figure 1) [15]). In passing, it should be briefly mentioned that the precise nature of the catalytic site of *Der p*1 appears to be ambiguous since work using cysteine protease- and serine protease-specific inhibitors have raised the possibility that the enzyme may possess a 'mixed cysteine-serine protease' activity, although more extensive work is needed to confirm this (Kalsheker, N., Q.M.C. Nottingham, personal communication). The observation that *Der p*1 exhibited homology to the cysteine proteases coupled with the knowledge of the allergen's presence at high concentration in the mite faeces implied that the allergen was a digestive enzyme in the gut of the dust mite, an observation which has had several implications for the allergen since then, as we shall now see: For example, in 1990, Herbert, *et al*, demonstrated that *Der p*1 could increase bronchial permeability to BSA (a 66kDa protein*) in vitro*, which thus implied that the mite allergen could broaden the range of molecules capable of penetrating the airway in an asthmatic patient, and hence increase the likelihood of challenge to the immune system, a factor which could potentially directly accentuate the damage caused to the airway during an asthma attack [16].

A second role for *Der p*1 in asthmatic patients was suggested by Hewitt, *et al*, in 1995 [17]. The team suggested that the allergen could selectively cleave the low-affinity IgE Fc receptor (CD23) from the surface of human B-cells. By doing this, *Der p*1 enhances IgE responses by removing an important feedback inhibitory mechanism that normally limits IgE synthesis [17]. Moreover, the soluble IgE receptor released from the cleavage event is thought to, itself, further promote IgE synthesis [17]. So in this way, once into the bloodstream, it is thought

that *Der p*1, through its catalytic activity, can greatly enhance IgE accumulation which may in turn lead to the severe inflammation associated with asthma.

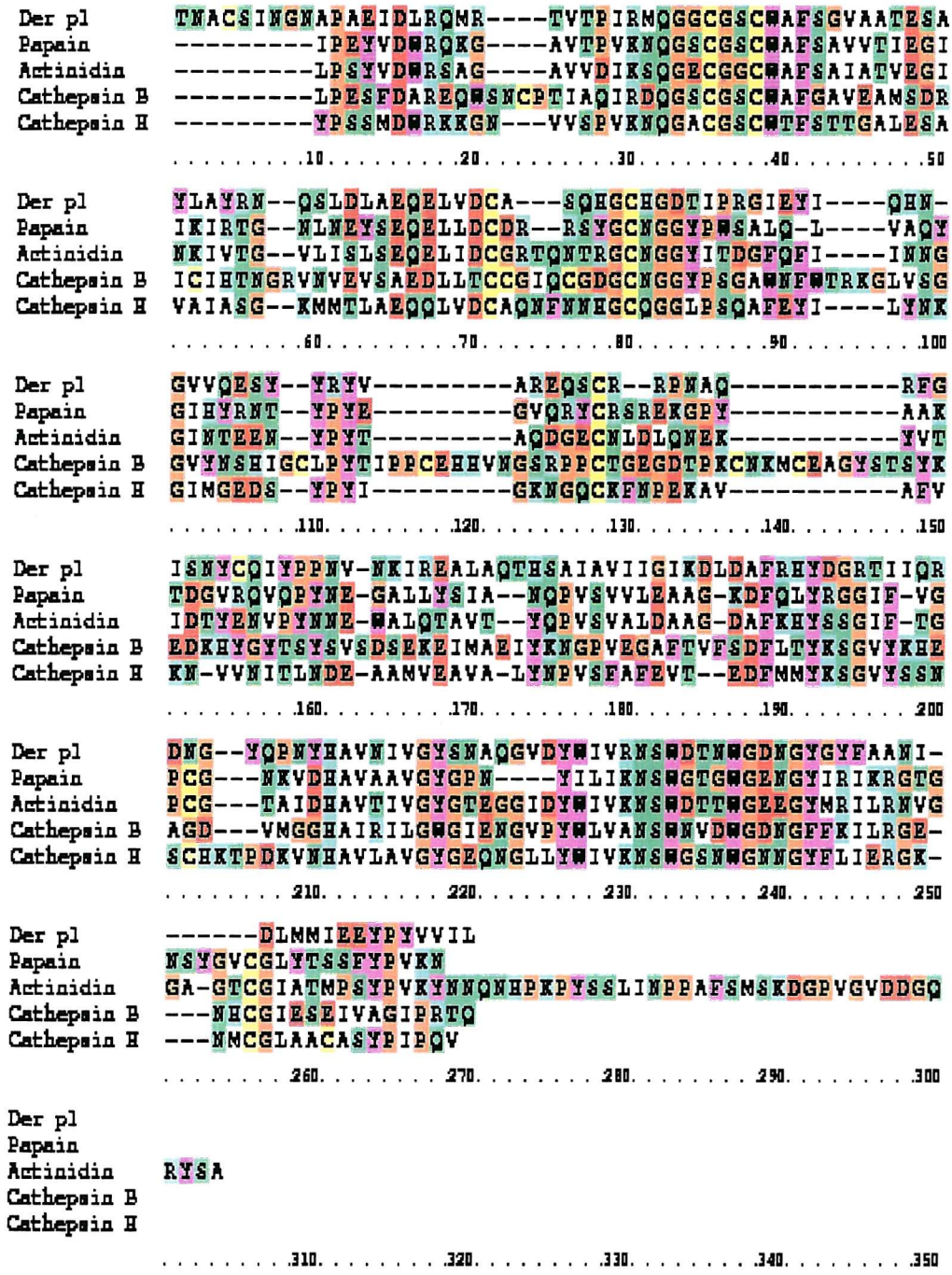


Figure 1. CINEMA (version 2.1) [18] alignment of mature 222 amino acid *Der p*1 sequence [14,19] with several members of the cysteine protease family. The alignment procedure was performed using MALIGN [20]: Percentage identities between *Der p*1 and each of the other homologues were: papaya papain [21], (28.8%), kiwi actinidin [22] (32.9%), rat cathepsin B [23] (22.5%) and rat cathepsin H [24] (32.3%). **Key**: Residue types - Blue= Polar positive, Red= Polar negative, Green= Polar neutral, Purple= Non-polar aromatic, White= Non-polar aliphatic, Brown= Proline/Glycine - the latter have special structural properties and so are grouped separately - similarly Yellow= Cysteine, and is classified separately due to its potential for disulphide bond formation. Dashes (-) denote padding characters. Sequences were obtained from the SWISS-PROT/TrEMBL protein sequence database [25].

A third mechanism by which *Der p*1 may be involved in asthma based on its catalytic potential, may lie in its apparent ability to catalytically-inactivate the airway protective agent α_1-antitrypsin, as reported by Kalsheker, *et al*, 1996 [26]. The latter is an important enzyme inhibitor which is highly effective against potentially-destructive proteases such as elastase [26]. Related to this, evidence has emerged based on the observation that atopic asthmatics recover from asthma after transplantation of lungs from non-asthmatic donors, which suggests that there is a particular primary defect localised in the bronchial epithelial wall of asthmatic patients which allows access of allergen into the body [27]. The hypothesis has been proposed that the nature of this defect may be pinned down to defective anti-proteases produced at the lung surface by the bronchial epithelium of asthmatic individuals, which are the primary line of defence against the inhaled allergen molecules. Without these anti-proteases, the allergen is free to bury its way into the body unchallenged. The evidence for this suggestion comes from studies involving *Der p*1 exposure to cells of normal and asthmatic patients, which have shown that after an hour, the allergen is removed from cells which it invades in healthy individuals, whereas it persists in the bronchial epithelia of asthmatics [28]. Moreover, a recent study has shown that the human lung surfactant proteins A (SP-A) and D (SP-D) which transfer microorganisms at the lung surface to phagocytes, can also bind *Der p*1 in a carbohydrate-specific and calcium-dependent manner. Furthermore, both proteins can inhibit allergen-specific IgE binding to mite extracts either via steric hindrance or competitive binding, and thus may be involved in the modulation of allergen sensitisation and/or the development of allergic reactions [29].

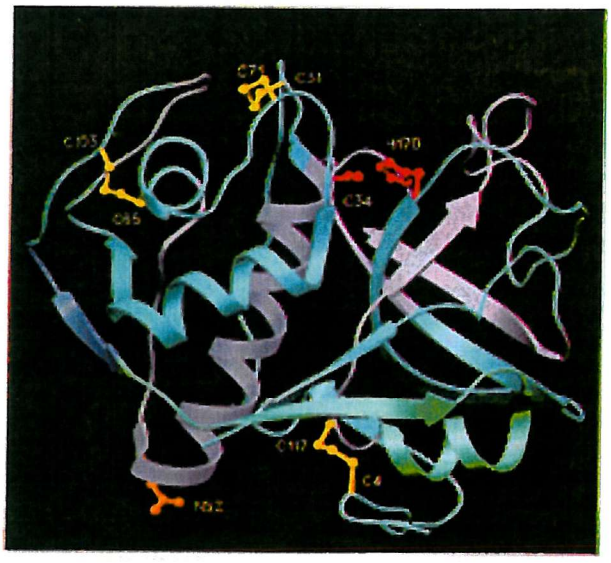
A novel suggested function for *Der p*1 in asthma has recently been postulated. This is its involvement in the upregulation of expression of several cytokines, including granulocyte/macrophage colony-stimulating factor (GM-CSF) [28], interleukin (IL)-8 and tumour necrosis factor α (TNF α), which serve to enhance the severity of the inflammatory

response by interacting with various immune system components (GM-CSF, for example, recruits macrophages with antigen-presenting phenotypes from monocytes, which increases autologous CD4(+) T-cell proliferation [30]). The mechanisms by which the allergen performs this role are unclear, but it seems that the latter, in some way, removes a cytoplasmic blocking protein, I*k*Bα, which allows access of NF-*k*B, a cytoplasmic transcription factor, to the nucleus which, in turn, induces expression of a number of important cytokines involved in the localised inflammatory response, including those mentioned above [31].

*Der p*1 has also, very recently, been shown to reduce the levels of interferon-γ in the blood, possibly indirectly via the enzyme's proteolytic activity, the result of which is an imbalance in the levels of the polarising cytokines interleukin-4 and interferon-γ which allows the allergen to further-modify the immune response [32].

Finally, a special mention should be made concerning the work of Topham, *et al* [33], who produced a cysteine protease fold-structural model for *Der p*1 based on its primary sequence homology with the known structures of three other cysteine proteases (viz. papain, actinidin and papaya proteinase Ω) (see figure 2). The structure was modelled using COMPOSER and the molecules were aligned by secondary structural elements as well as homology.

(a)



(b)

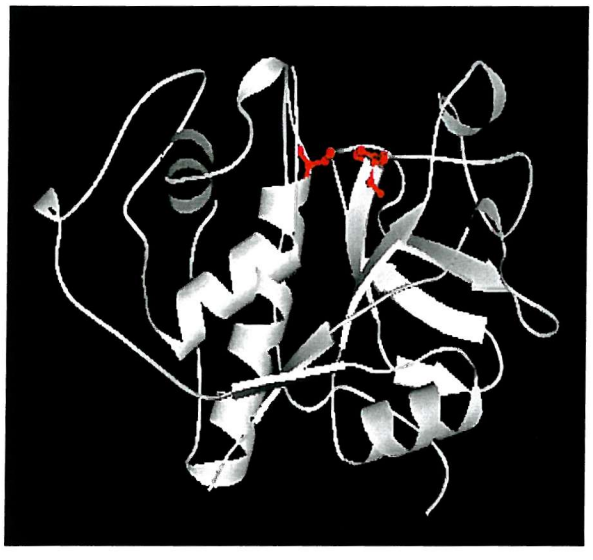


Figure 2. (a) SETOR [34] -generated molecular model of *Der p*1 as published by Topham, *et al* [33]. The three disulphide bonds are coloured yellow, the active site Cys-34/His-170 ion pair, red, and the glycosylation site(Asn-52), orange. Human B-cell epitope sites (residue ranges 1-33, 60-94, 101-106, 155-187 and 209-222) are coloured light blue, whilst T-cell epitopes (residues 112-131) are coloured green. A small region (residues 107-111) is epitopic for both B- and T-cells, and is coloured violet, whilst the remainder of the molecule is indigo. (b) MOLSCRIPT [35] representation of the three-dimensional structure of papain (1.6Å resolution) for comparison with the molecular model of *Der p*1. The coordinates used here were released by Pickersgill, *et al* [36], and were downloaded from the Protein Data Bank [37]. (It should be noted that this structure contains an additional atom bound to the active site Cys-25). With both molecules, arrows indicate β-strands, and coils, α-helices.

1.2 Aims of the work with *Der p*1

The overall aim of this work is to produce a three-dimensional structure of *Der p*1 which would allow detailed analysis of the active site with the goal of identifying special features that might be useful in the designing of specific inhibitors of enzymic activity. These could thence, for example, be applied to carpets/upholstered furniture to inactivate the allergens. The three-dimensional structure of *Der p*1 may also allow the identification of antigenic surface groups, specifically discontinuous epitopes, which would guide the design of mutant allergen for immunotherapy - i.e. to suppress the allergic immune response. This is particularly important in the light of research indicating that the only specific IgE binding epitopes present on *Der p*1 and *Der f*1 appear to be discontinuous in nature [38].

1.3 Momordin and RIPs

The Ribosome-Inactivating Proteins (RIPs) are a group of toxins found in relative abundance throughout the plant kingdom, and, less frequently, in bacteria [39,40,41,42] and fungi [43], which are capable of enzymatically preventing cellular translation. The RIPs are subdivided into two distinct classes: The Type I- and Type II- toxins [44,45,46].

The **Type I** toxins consist of a single polypeptide chain typically of about 26-31kDa and have characteristic pI values of between 8 and about 10. Examples of these toxins include trichosanthin and momordin (also known as momordin I or α-momorcharin) [47].

The **Type II** toxins are heterodimeric proteins, the two subunits being linked together by a disulphide bond. The **A-chain** of these Type II molecules shares homology with the single polypeptide chain of the Type I toxins, and as such, possesses the ribosome inactivating function of the toxin. The **B-chain** in comparison is a lectin involved in cell attachment and facilitates A-chain entry into the cytoplasm. Examples of the Type II toxins include ricin and abrin [47].

In addition to the two main classes of RIPs, recent work has uncovered a novel group of **small RIPs**. These include γ-momorcharin (11.5kDa) from the seeds of *Momordica charantia* [48] and LuffinS (1), LuffinS (2) and LuffinS (3) (all about 8kDa) from the seeds of *Luffa cylindrica* [49]. Work with γ-momorcharin has revealed that the substrate specificity is identical to that displayed by other RIPs belonging to either class I or II [48].

Additional novel RIPs which do not fit into either the class I or II categories have been identified and include an RIP from maize which is a two-chain toxin composed of 16.5- and 8.5kDa subunits which are non-covalently associated [50] and the bacterial Shiga toxin from *Shigella dysenteriae* serotype 1 and Shiga-like toxins I and II from certain *Escherichia coli* (*E. coli*) strains which all share high structural identity and consist of an approximately 30kDa active subunit which is associated non-covalently with a pentamer of 7.7kDa B subunits which are able to bind the membrane glycolipid, globotriaosylceramide [51,52].

The potency of the Type II RIPs is clearly evident when you consider that for the Type II RIP ricin, the K_M for ribosomes = 0.1μM, and the k_{cat} = 1500 min^{-1} [53]. As a consequence, organisms synthesising the toxins have evolved mechanisms to prevent damage to their own ribosomes. Various mechanisms exist, and one of the most studied has been that for ricin: The protein is synthesised as a preproprotein with a 12-residue linker bridging the gap between the two ricin subunits, and a 35 residue N-terminal sequence, part of which represents a signal peptide [54]. The preproprotein is inactive towards ribosomes [55]. The zymogen is co-translationally translocated into the endoplasmic reticulum (ER) where the signal peptide is removed [56]. The toxin then undergoes *N*-glycosylation [57] and disulphide rearrangement [58] during its time in the ER and is then passed through the Golgi network and finally targeted to the protein bodies [57], possibly via its internal 12 amino acid linker peptide [59], which subsequently, along with the N-terminal propeptide, is cleaved by a protease [60,61] and the toxin is thence stored in its active heterodimeric form in the organelle matrix, from which it apparently cannot escape [62].

In general, plants appear to get round the problem of self-ribosome-inactivation as illustrated above, by targeting their RIPs, via the ER/Golgi network, to different cell compartments such as the protein bodies, vacuole [63,64,65], cell wall matrix [65] and the region between the

primary cell wall and the plasmalemma [64]. RIPs from cereals such as barley and maize however, appear to have no signal peptides and so are thought to reside in the cytosol of cells [50,66]. It has been shown that the ribosomes of these plants are resistant to the action of their endogenous endosperm RIPs [67,68]. It should be noted that some of these cereal RIPs may also be present in the cytosol in an inactive precursor form [50,69].

The next question we should try to address is how the RIPs gain access to the cytosolic ribosomes in 'target' cells in order to exert their cytotoxicity. It should be noted here that the type II RIPs are far more potent in this regard than the type I toxins. This is simply a consequence of the former's lectin B-chain which readily allows access to cells via receptor molecules using its galactose/*N*-acetylgalactosamine binding properties. Ricin has once again been the most carefully studied toxin in this area of RIP research, and as such it is ricin that we will use as our model to describe the postulated mechanisms of RIP entry into the cytosol. Toxin uptake occurs via the formation of both clathrin-coated vesicles and non-clathrin coated vesicles which arise through a dynamin-independent process [70]. Uptake of ricin by either route results in toxin transfer to endosomes. From here, electron microscopy has revealed that the RIP has several fates including being recycled back to the cell surface and sent off to the lysosomes, but importantly, a small, yet significant proportion accumulates in the trans Golgi network [71]. At this point, experiments revealing that the ricin A-chain becomes core glycosylated at some time during its journey to the cytosol, a feature diagnostic of passage through the ER (the home of the enzyme oligosaccharyl transferase) have suggested that the toxin follows a retrograde transport pathway to the ER [72]. From here, however, it is uncertain how the RIP accesses the cytosol. The currently-favoured hypothesis [73] is that the toxin may, through appearing as a misfolded protein, take advantage of a pre-existing membrane transport system such as that used for exporting damaged proteins of the

ER to the cytosol for degradation [74,75]. The details of this ER to cytosol transport system are still unclear, but the proteinaceous Sec61p channels that are associated with the import of nascent polypeptides from the cytosol appear to be involved [76,77,78]. Once out into the cytosol, the RIP must overcome one final hurdle if it is to survive to reach the ribosomes, this is the ubiquitin-proteasome degradation pathway which targets and degrades aberrant proteins exported to the cytosol [74,76]. Ricin has a deficiency of lysine residues [79] and is also highly protease-resistant [80], two properties which make the toxin resistant to ubiquitination and degradation respectively. For these reasons the toxin may be able to evade the destructive pathway for long enough to fulfil its role and inactivate the ribosomes.

Workers elsewhere have proposed an alternative, more direct pathway of RIP entry into cells: they have postulated that the toxin can physically interact with membrane lipids and in doing so can induce vesicle-vesicle fusion. In this way, it is thought that the toxins are able to gain passage to the cytosol from the vesicular trafficking pathways by a mechanism which involves the disruption of vesicle membrane integrity upon membrane-membrane fusion [81]. Different RIPs have shown different properties in being able to induce membrane fusion: ricin and viscumin act as direct fusogens, whilst *Ricinus communis* agglutinin operates as a non-fusogenic toxin which is able to alter fusion between artificial membranes after modifying intramembrane contacts [81]. It was shown, however, that the ability of RIPs to induce fusion was critically-dependent on the composition of the membranes concerned, small variations resulting in no observable fusogenicity. On the other hand, it has been shown that β-galactoside-specific lectins, upon interacting with certain cell surface glycoproteins which have signalling potential, are capable of promoting fusion by elevating the levels of phospholipids and inositol phosphates [82], i.e. diacylglycerol activation leads to improved liposome fusion [83]. Although at first glance it may appear that the retrograde transport method for RIPs is the most experimentally-well defined and thus favoured potential pathway

for getting RIPs into cells, it should be noted that when certain cells are treated with brefeldin A (a Golgi apparatus disrupting agent) [84], and when certain aberrant Golgi complex-expressing cell lines are used [85,86], their ribosomes have been shown to still be inactivated by endocytosed RIPs, which are observations incompatible with the retrograde transport toxin hypothesis, but which do not conflict with the vesicle fusion hypothesis, thus there may be more than one pathway of entry into the cytosol of target cells by RIPs.

The biological function of these toxins is uncertain, though it's thought that plants may produce the RIPs as a self-defence mechanism against insect- [87], virus- and fungal attack or alternatively as a means of eliminating altered ribosomes [88].

So, why study these RIPs? The toxins have potential therapeutic application in cancer and in AIDS [89,90]. RIPs could, for instance, be coupled to antibodies, allowing specific delivery to tumour cells [89,91], and it is this type of thinking that has led to an intense focus of research into RIPs being directed towards combating cancer through various conjugations of toxins to targeting molecules. To this end, **immunotoxins**, as just discussed, are currently the preferred construct used, although novel alternatives are occasionally seen, such as **hormonotoxins**, for example RIPs coupled to the luteinising hormone, lutropin [92], which target membrane receptors in tumour cells in order to deliver their toxins, **RIP-growth factor constructs** such as the RIP-heparin-binding epidermal growth factor fusion proteins which target numerous tumour cell-types over-expressing the epidermal growth factor receptor [93], and finally, **RIP-ion transfer protein constructs**, such as the RIP-transferrin conjugate which again targets a wide range of tumour cells due to their over-expression of transferrin receptors [94]. Moreover, some of these studies have shown that the potency of these chimeric toxins can be enhanced by the addition of certain drugs such as monensin and

chloroquine, which raise intracellular pH, and in doing so, it is thought, influence toxin stability and activation [94,92].

Single chain RIPs have also been shown to selectively inhibit human immunodeficiency virus (HIV) replication in infected cells at concentrations an order of magnitude lower than those levels required to kill human cells [90]. This suggests they could be used therapeutically to inhibit HIV replication in the cell. Indeed, as a further example of this potential, work has been carried out using an RIP immunotoxin (TXU (anti-CD7)-pokeweed antiviral protein conjugate) as an anti-HIV type 1 agent which has displayed potent anti-HIV activity in a mouse model of human AIDS without causing any side effects and at doses that were very well tolerated by cynomolgus monkeys. Moreover, plasma samples from TXU-PAP-treated cynomolgus monkeys displayed potent anti-HIV-1 activity *in vitro* [95].

In passing it should be quickly mentioned that it has long been known that RIPs display abortifacient activities. Indeed, the root tubers of *Trichosanthes kirilowii* have been used for centuries in Chinese herbal medicine to induce mid-term abortion [96]. The mechanisms by which trichosanthin functions as an abortifacient agent have been extensively studied [97]. The RIP (a) brings about the selective death of placental villi syncytiotrophoblasts; (b) this results in the formation of clots in the local circulation which leads to large areas of infarction; (c) this, in turn, leads to a drop in human chorionic gonadotropin [98] and steroid hormone levels, a disruption of metabolic exchanges and a rise in prostaglandin synthesis with resultant onset of abortion [99].

Both the type I and -II RIPs possess specific *N*-glycosidase activity which is the key to their toxicity [100,101]. Both function by inactivating ribosomes in precisely the same way, and the absolute specificity and simplicity of this inactivation process is astonishing! Both RIP classes recognise a striking ribosomal RNA (rRNA) region termed the sarcin/ricin loop (SRL)

(see figure 3 below) which contains many universally-conserved nucleotides [103] and which is known to be essential for elongation factor binding to the ribosome (elongation factors 1 and 2 in eukaryotes [104,105] and elongation factors Tu and G in *E. coli* [106,107]).

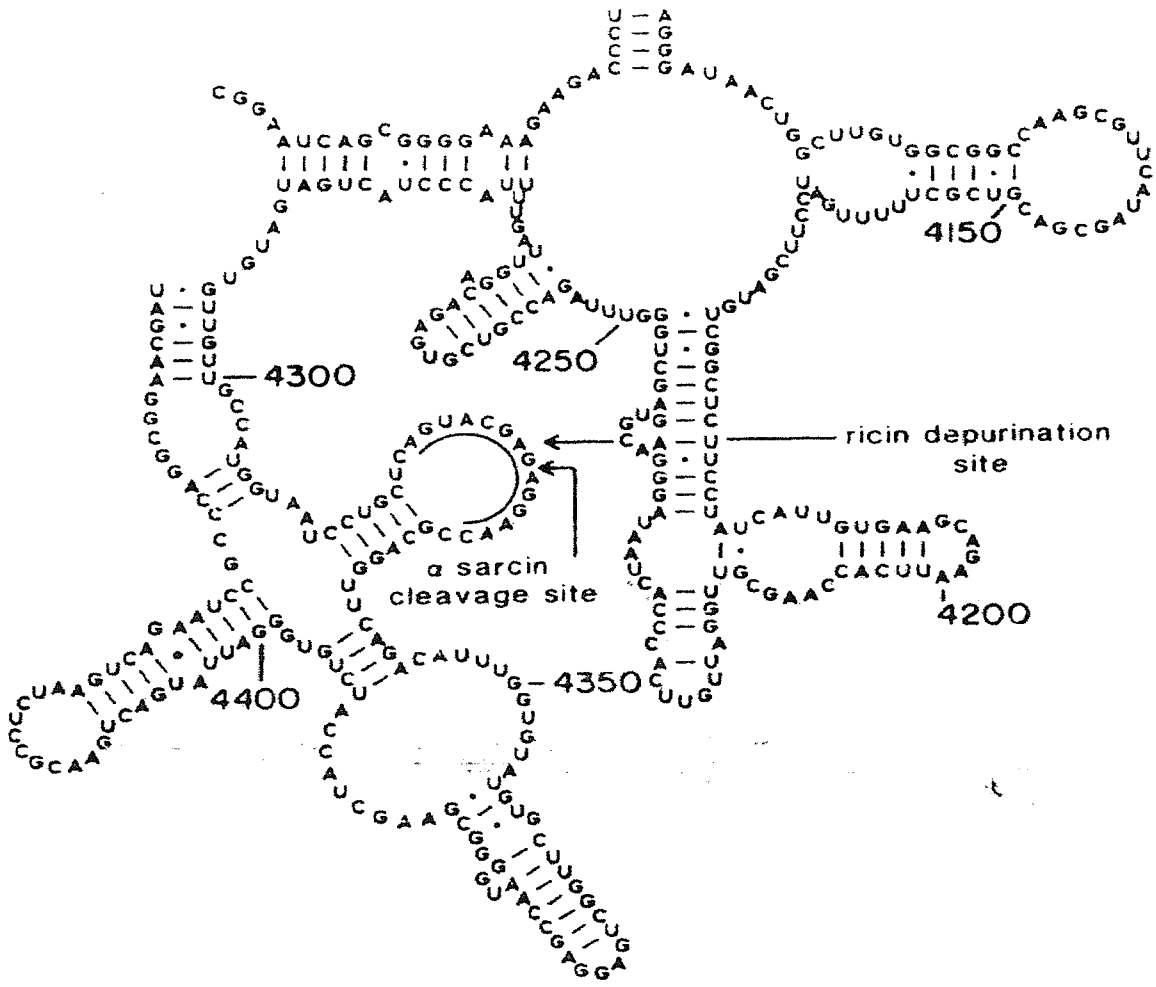


Figure 3. Simplified two-dimensional representation of the region of eukaryotic 28S rRNA containing the sarcin-ricin loop (nucleotides 4312-4338). Both the ricin and α-sarcin cleavage sites are indicated with arrows. Near-universal nucleotides are underlined, a "-" indicates a canonical Watson-Crick base pair, whilst a "." represents a non-canonical Watson-Crick base pair. The adenosine attacked by ricin here is designated A4325 which is equivalent to the A4324 mentioned in the main body of the text, this is due to some confusion in the literature over the numbering of the nucleotide sequence. This figure has been reproduced from the paper by Glück, A., Endo, Y. and Wool, I.G., 1992 [102].

The SRL is so-called since it was identified as being a substrate for both ricin [100,101] and another translation-halting enzyme, the ribonuclease α-sarcin [108,103]. It is found at only one position in both eukaryotic 28S rRNA [100,101] and in prokaryotic 23S rRNA [109] and its basic structure is a double-helical stem at the base of a hairpin 'loop' containing 17 and 15 nucleotides respectively [109,110] (It should be noted that ricin and other RIPs will also target an additional region in prokaryotic 16S rRNA where there is just a 4-nucleotide 'loop' present [109,102], and also that in contrast to the eukaryotic SRL, cleavage of the prokaryotic 'loops' will only occur with **naked** rRNA [100,109]).

At the top of the SRL (position A4324 in rat 28S rRNA [101] and position A2660 in *E. coli* 23S rRNA [109], and also in position A1014 in the tetraloop found in *E. coli* 16S rRNA [109]), there is a critically-conserved adenosine whose adenine-ribose glycosidic bond has been identified as the absolutely-specific site of hydrolysis by ricin and all other RIPs, i.e. it is this single cleavage event which results in the complete inactivation of the ribosome.

Mutation studies with synthetic oligoribonucleotides mimicking the rat 28S rRNA SRL [110] and the ricin-cleavable *E. coli* 16S rRNA tetraloop [102] have unveiled ricin's catalytic-requirement for (i) an intact helical stem and (ii) the presence of a 5' GAGA 3' nucleotide sequence in a precise position at the top of the 'loop' in the substrate rRNA, where the first adenine in the GAGA sequence is the aforementioned RIP-cleavable base.

A detailed structure of the SRL has now been obtained from early nuclear magnetic resonance (NMR) spectroscopy- [111,112] and recent X-ray crystallography [113] work with a synthetic 29-nucleotide RNA that has the SRL sequence. The NMR-determined structure revealed for the first time additional, and extensive, sub-structure within the 'loop' region of the SRL which could be partially-attributable to non-Watson-Crick base pairings. It also uncovered a bulged-guanosine motif which is involved in the ribonuclease α-sarcin-SRL interaction, and an exposed tetraloop structure for the aforementioned GAGA nucleotide sequence where the

flanking guanosine and adenosine in the sequence form a non-Watson-Crick base pair forcing out the RIP-sensitive adenosine into a very exposed position in a dinucleotide loop. Finally, the NMR-derived structure also confirmed the 'loop' as having an A-type double-helix at its base.

This work was followed up several years later by the determination of the X-ray crystallography structure at 2.1Å resolution, which provided a more detailed picture of the secondary structure of the SRL: the 'loop' structure was divided up into clearly-defined motifs, a "flexible region" comprising three non-Watson-Crick base pairs which is found just above the helical stem, a "G-bulged cross-strand A stack" which lies above that, and finally, the aforementioned "GAGA tetraloop" which sits right at the very top. The X-ray data also revealed an "S-turn" motif where the backbone adjacent to the bulged guanosine (G4319 in rat 28S rRNA) is distorted by two sharp bends giving it an S-shaped appearance.

As a brief aside, the NMR and X-ray structures of the synthetic SRL are amongst the first RNA structures ever elucidated by **both** techniques and have subsequently provided valuable information validating the application of NMR - a methodology initially developed for the determination of protein structures - to RNA [114].

It should be briefly mentioned that the precise GAGA loop structure must be rigidly fixed for ricin action, but that it is not so critical for all RIPs, e.g. pokeweed antiviral protein will still cleave off the specific adenine if the highlighted guanine (GA**G**A) is mutated to adenine [115].

The question of substrate requirements for RIPs has been further complicated by work which has demonstrated RIP-induced adenine release from mRNA, tRNA and DNA in addition to the ubiquitously-observed activity against rRNA. It was suggested therefore from this evidence that the RIPs may have the potential for an even larger involvement in halting cellular translation [116].

A small group of RIPs are cofactor-dependent, requiring ATP and tRNA to reach maximal activity on isolated ribosomes. Specific tRNA molecules are required for each RIP and these are currently becoming unveiled [117].

It should be noted in passing that some RIPs have demonstrated additional catalytic properties, such as DNase activity [118] (where specific undamaged, non-mispaired adenines are first removed from a single-stranded substrate which then leads to cleavage at the resultant abasic sites), RNase activity [119,120] (several RIPs have demonstrated such activities on rRNA and α-/β-momorcharins have also shown such activities on tRNA, at pH values below 5.5. Different RIPs also possess different base specificities on RNA molecules) and chitinase activity [121] (three plant RIPs have been shown to exhibit endochitolytic activity in addition to their specific 28S rRNA *N*-glycosidase activity). The biological significance of these DNase/RNase activities is unclear, but the chitinase activities of the plant RIPs may provide a competitive advantage (e.g. improved antifungal protection) to the host plants.

Crystallographic studies on the Type II RIP ricin have revealed many structural details, for instance, it is known that the A-chain is a globular protein with extensive secondary structure (30% α-helix, 15% β-sheet) [122]. This subunit seems to be composed of three somewhat arbitrary domains: a five-stranded β-sheet (residues 1-117), a cluster of α-helices (residues 118-210) and an additional compact region at the C-terminal end (residues 211-267) [122]. The active site of the ricin A-chain is located at the interface between all three domains and of the active site residues, it is Arg-180, Glu-177, Asp-96 and Tyr-80 which are thought to be especially important for enzymic activity, as we shall see shortly [122,123,124]. Table 2 shows a list of active site residues in a number of different RIPs and also displays conserved residues in all RIPs [124].

	14	17	22	70	83	85	108		114	155	160	163	189	192
			* * *	*	*	* * * *	*		*	* *	*	*	*	
TCS	Y G V	F...L	R...V	Y I...F	N -	E...S	G N Y E	R	L...I	Q...S	E A A	R...E	N S	W
MMC	* G M	*...L	*...I	* I...F	N -	E...S	G N Y E	R	L...I	Q...A	* A A	*...E	N S	*
RTA	* T N	*...V	*...A	* V...F	H P	D...G	G N Y D	R	L...I	Q...S	* A A	*...E	N S	*
SAP	* S S	*...I	*...L	* V...F	R S	E...E	Y T+Y D	S	I...I	Q...A	* V A	*...E	V S	*
PAP	* A T	*...L	*...L	* V...F	N -	D...N	G L Y P	T	L...I	Q...S	* A A	*...E	N -	*
LUF	* S E	*...L	*...L	* I...F	N -	E...S	G N Y E	K	L...L	Q...A	* A S	*...E	N S+	*
ABR	* K Q	*...L	*...A	* V...L	R -	D...Y	G T Y G	D	L...I	Q...A	* A A	*...E	N N	*
BPI	* A T	*...I	*...I	* L...W	W -	E...-	G D T D	K	L...L	-...N	* A T	*...Q	+N G	*

Table 2. A list of conserved amino acids and those representing the active site region of a number of RIPs. Numbers represent either universally-conserved amino acids or the starting residue in a sequence region present in the active sites of the RIPs trichosanthin (TCS) and momordin (MMC). TCS is the source of the numbering. Asterisks above the sequences represent residues present at the active site. Asterisks within the sequences themselves show universally-conserved residues. + denotes insertion of residue(s) whilst - represents deletion of residue(s). RTA is ricin A-chain, SAP: saporin SO-6, PAP: pokeweed antiviral protein, LUF: luffin-α, ABR: abrin A-chain and BPI: barley protein-synthesis inhibitor. For clarity, Tyr-70, Glu-85, Glu-160 and Arg-163 in trichosanthin are equivalent to Tyr-80, Asp-96, Glu-177 and Arg-180 in ricin A-chain. (Table taken from Huang, *et al* [124]).

Crystallographic studies on several RIPs have suggested alternative catalytic mechanisms for the ribosome inactivation process. Three such sets of studies include the work on the Type II RIP ricin A-chain (Robertus, *et al* [125]), the work carried out on the Type I RIP α-momorcharin (an alternative name for momordin) (Ren, *et al* [47]) and the research involving both trichosanthin (TCS) and α-momorcharin (Huang, *et al* [124]).

The work of Robertus *et al* involved the determination of crystal structures of substrate analogue formycin 5'-monophosphate(FMP)- (2.8Å resolution) and adenyl (3'→5')guanosine (3Å resolution)-ricin complexes. These studies allowed the team to propose a mechanism of RIP action which began with ricin's Arg-180 directly transferring a proton onto the N3 position of the target adenine, which then, after electron movements, results in cleavage of the sugar-base glycosidic bond and the formation of a sugar

oxycarbonium ion. The latter is stabilised by a local glutamate (Glu-177). The stabilising addition of a water molecule completes the reaction whose hydroxyl stabilises the sugar and whose remaining proton is donated to the RIP Arg-180 to yield the reaction products (see figure 4 below for a diagrammatic summary of this proposed mechanism) [125].

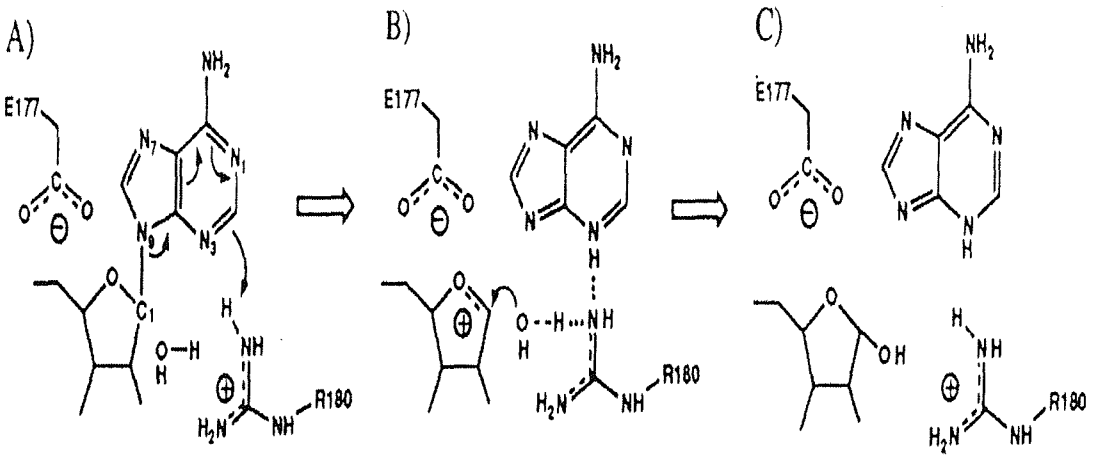


Figure 4. A diagrammatic representation of a proposed mechanism of ricin action. See text for details. A) and B) give details of the bond breaking between the ribose and adenine, whilst C) is concerned with bond formation to the attacking water. Note: a proton is **directly transferred** by Arg-180 to adenine's N_3 in order to initiate the reaction. Illustration taken from *Monzingo and Robertus, 1992* [125].

The work of Ren *et al* provided high resolution crystal structures for native, product- and substrate analogue FMP-complexed α-momorcharin. The mechanism postulated from these studies again involved the RIP's Arg-163 interacting with the target adenine at the N3 position. This time, however, instead of direct protonation occurring, a partial positive charge arises within the base which then spreads across it and which, in combination with postulated conformation-induced strain about the ribose-adenine glycosidic bond, leads, via stabilising electron shifts, to cleavage of the latter. Glu-160 from α-momorcharin can then stabilise the

resultant ribose oxycarbonium ion and a water molecule finally associates with and stabilises the sugar and organic base to yield the products (see figure 5 below) [47].

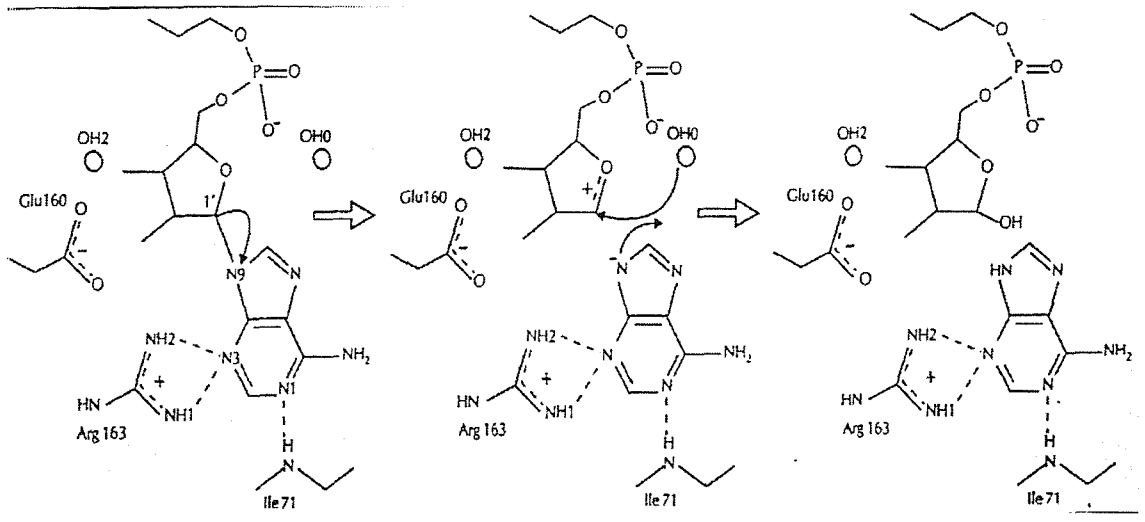


Figure 5. Diagrammatic representation of Ren, *et al*'s proposed mechanism of action of momordin. See text for details. Note that in contrast to the work of Monzingo and Robertus, Arg-180 only partially-protonates the adenine's N_3 here. Illustration taken from Ren, *et al* [47].

Huang, *et al*, have performed structural work on complexes between both adenosine triphosphate (ATP) and formycin and the RIP α-momorcharin (1.6-2.0Å resolution) and have also collected data on native α-momorcharin alone (2.2Å resolution). From these studies they have proposed the following mechanism:

The susceptible nucleoside first binds to the so-called "second part" of the RIP's active site where direct protonation of the adenine's N-7 occurs via an acidic active site residue (Glu-85 in TCS and α-momorcharin) whose side chain is held in precise position by the nearby Arg-163. Next, the Tyr-70 side chain dramatically rotates by 57° around its Cα-Cβ bond which opens up a pathway between both the first and second parts of the active site. The adenine then becomes trapped in the so-called "first part" of the active site at which point the glycosidic bond begins to break as electrons move from here to stabilise the positively-charged base. The resultant oxycarbonium ion is stabilised by the RIP's Glu-160 which also

activates a local water molecule that hydroxylates the sugar and thus neutralises its net charge. Finally, the free adenine leaves the active site and the Tyr-70 rotates back to its initial position, closing off the two active site chambers and thus completing the reaction (see figure 6 for an illustration of this mechanism) [124].

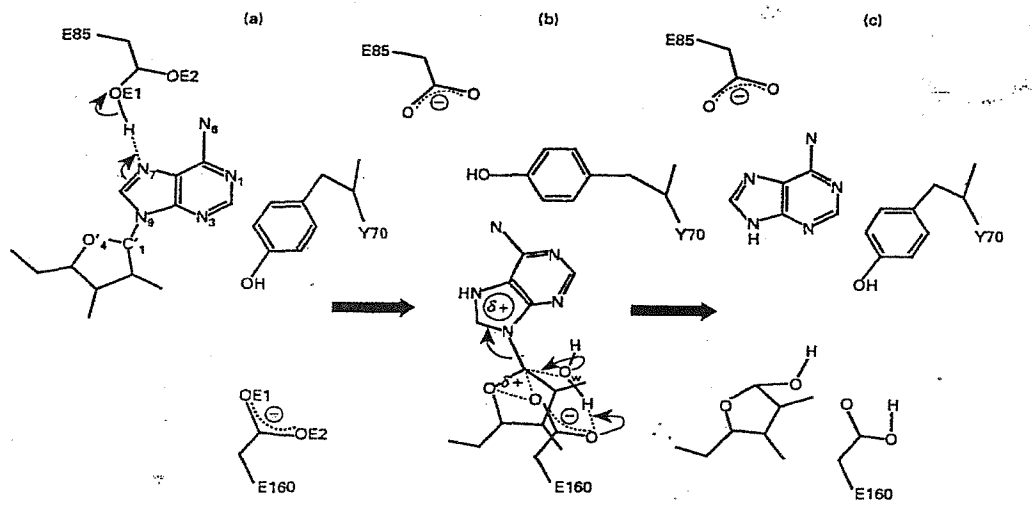


Figure 6. Illustration of a possible mechanism of RIP action as proposed by Huang, *et al.* See text for full details. (a) Protonation of the cleavable adenine via Glu-85. (b) Involvement of Glu-160 in oxycarbonium ion stabilisation and bond cleavage. (c) Reaction products. Notice how Tyr-70 shifts position to allow substrate/product movement in/out of the active site. Also note that in contrast to the other proposed mechanisms, the initial cationic attack on the adenine here proceeds via N_7. Taken from Huang, *et al* [124].

Recent work on the RIPs has thrown up lots of new intriguing data: The structure of the Type I RIP trichosanthin bound to NADP has now been solved where, surprisingly, the bound nucleotide remained intact [126]. This work led to a development of the interest of our group (Husain, *et al*) in momordin, a Type I RIP from the seeds of the bitter gourd, *Momordica charantia*, whose structure the group had previously determined [127]. Crystals of momordin and $NADP^+$ were grown and revealed a very interesting finding: the nicotinamide group had been cleaved in preference to the adenine. The cleaved nicotinamide stayed in the active site of momordin and could not be removed with ease, which potentially has implications concerning the mechanism of RIP action.

1.4 Aims of the work with momordin

The main aims of this work are to study the interactions between momordin and a series of several other compounds such as NADH/NAD^+ at the active site with a view to uncovering the nature of the mechanism by which momordin functions.

1.5 MhpC

Many bacteria have displayed an ability to degrade a number of phenylpropanoid compounds such as cinnamic acid, phenylpropionic acid and their hydroxylated derivatives [128]. These bacteria include *Acinetobacter* sp. (formerly *Achromobacter* sp.) [129], *Pseudomonas* sp. [130,131,132], *Arthrobacter* sp. [133] and *E. coli* [134,135]. This is significant since these enzymic activities allow the breaking of carbon-carbon bonds, which represent rare catalytic events, and which are of invaluable use in the degradation of several ubiquitous environmental pollutants (which ordinarily have a very slow rate of decomposition) to non-toxic products [136]. It is interesting to discover that certain *E. coli* strains can grow using particular phenylpropanoid compounds as sources of carbon [135]. This shows that these pathways of degradation can be important for generation of energy within the cell, which constitutes a biological purpose for these routes, thus explaining their existence. As we shall see later on, several products of these pathways of catabolism may be directly fed into the Krebs' cycle.

A means by which the bacterial enzymes of aromatic catabolism encounter their natural substrates, such as 3-phenylpropionic and 3-(3-hydroxyphenyl)propionic acids (3-HPP), has been described for *E. coli*, where these aromatics are released as products of reactions of parasitic bacteria in the animal gut, and derive from plant precursor molecules such as ferulic and caffeic acids, the flavinoids catechin, myricetin, and hesperetin, and from the amino acids tyrosine, phenylalanine and some of their metabolites [135]. Once the substrates are encountered, aerobic catabolism may then proceed close to the gut epithelial walls where oxygen diffuses away from the blood. Aerobic catabolism may also occur in soil, sediment and in water once *E. coli* is excreted from the intestines [137].

The bacterial degradation of aromatic substrates relies on several different enzymes during the initial steps of catabolism since there are many different primary substrates which feed

into these pathways, but the products of the primary substrates tend to be narrowed down to just a very few key dihydroxylated molecules, which are subsequently degraded to the aforementioned Krebs' cycle intermediates by two distinct groups of enzymes, those of the *ortho-* and *meta*-cleavage pathways [138].

The pathway of aromatic acid catabolism mainly studied in *E. coli* is that involving the Mhp enzymes from the *mhp* operon (see figure 7). These enzymes (MhpA-F) are responsible for converting 3-HPP to various metabolites as shown in figure 7 B)[128]. (It is from this ***meta***-cleavage catabolic pathway of 3-**HPP** in *E. coli*, that the abbreviation **Mhp** is obtained).

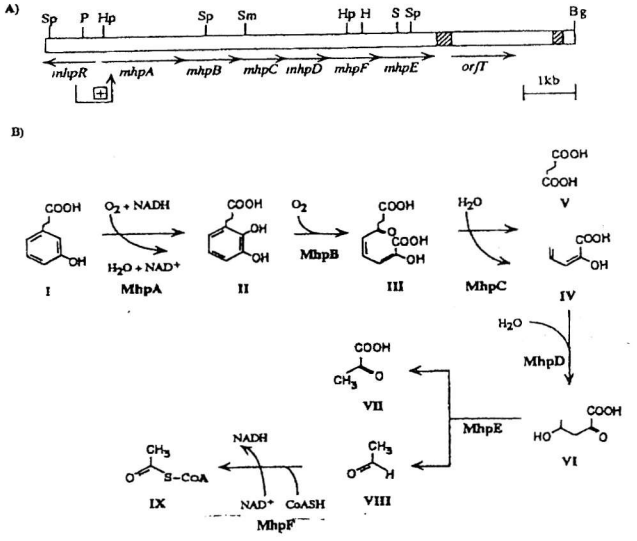


Figure 7. The Mhp system in *E. coli*: A) Arrangement of genes at the *mhp* operon. Certain restriction endonuclease sites are shown in relation to the genes, i.e., *Bgl*II (Bg), *Hind*III (H), *Hpa*I (Hp), *Pst*I (P), *Sac*I (S), *Sma*I (Sm), and *Sph*I (Sp). Arrows show the directions of gene transcription. In the presence of 3-HPP, the *mhpR* gene product enhances the expression of all the other genes. B) The pathway of 3-HPP catabolism. The genes *mhpA,B,C,D,E* and *F* correspond to the enzymes shown in this diagram. (Figure taken from reference 128).

Of particular note in this series of enzymes are MhpB, a non-haem-iron(II)-dependent oxygenase, responsible for converting 3-(2,3-dihydroxyphenyl)propionate to 2-hydroxy-6-ketonona-2,4-diene-1,9-dioate, and MhpC, a hydrolase which uses the latter as substrate, and converts it to succinate and 2-hydroxy-penta-2,4-dienoate [128]. MhpB has been shown to belong to a novel family of extradiol dioxygenases [128] and so has been of interest to this end. The attention on MhpC has been mainly mechanism-based. It is thought that the

homodimeric enzyme catalyses the reaction via an initial enol/keto tautomerisation. This then gives rise to an additional ketone group off of the substrate's carbon-2 which is proposed to act as an electron sink allowing the carbon-carbon cleavage reaction. Evidence for this enol-keto tautomerisation has derived from radiolabelling experiments in which incubation of MhpC along with substrate and heavy water (^2H_2O) yielded deuteriated product in the form of 2-hydroxypenta-2,4-dienoate, radiolabelled with deuterium at the C-5 position [139]. Further evidence derived from kinetic analysis of the MhpC reaction, which could only be explained by using a branched kinetic mechanism in which the intermediate is released at a rate comparable to its catalytic turnover [140]. Both pieces of evidence are consistent with the proposed enzyme mechanism. The suggested mechanism can be seen in figure 8 [141].

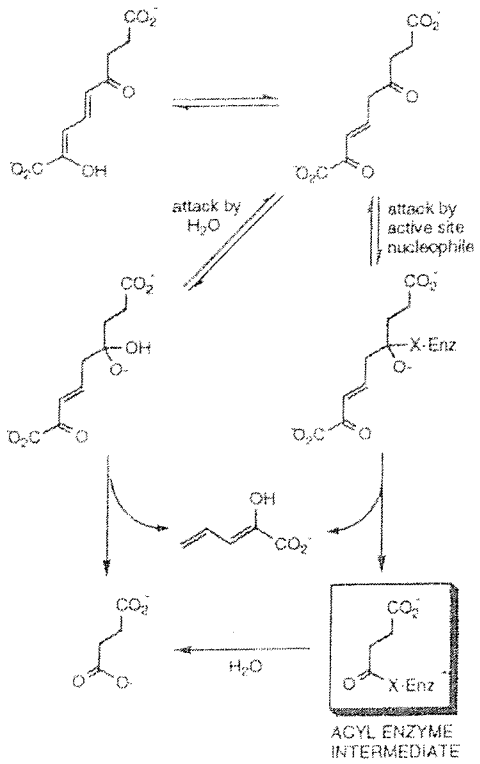


Figure 8. Two alternative mechanisms proposed for the MhpC-catalysed reaction. Note the initial formation of a C-2 ketone intermediate which serves as an electron sink, allowing cleavage of the carbon-carbon bond to proceed. Attack of the C6 ketone group is achieved either by an activated water molecule or by an active site nucleophile. Both putative mechanisms are shown (Figure taken from reference [141]).

It can be seen from this mechanism that a nucleophile is required to attack the C-6 carbonyl carbon. This may be an activated water molecule as seen with aspartic proteases [142] or it could be an amino acid side chain. Recent radiolabelling experiments have suggested that the nucleophile is an activated water molecule [141]. This is because no evidence to indicate the presence of an acyl-enzyme intermediate has been found, which would exist if an amino acid served as a nucleophile. However, the possibility remains that an active site nucleophile is involved in the MhpC mechanism. The most-likely candidate for this is Ser-135. The evidence for this comes from homology data (unpublished, see figure 9) which has suggested that MhpC belongs to a family of enzymes possessing the α/β hydrolase fold, a structural motif where a three-dimensional core structure (an α/β-sheet, not barrel, of eight β-strands connected by α-helices) is conserved despite the absence of significant primary structural similarity [143,144]. Enzymes with this fold have dramatically diverse catalytic specificities and yet all appear to share similar reaction mechanisms [143,144]. Sequence alignments, chemical modification and mutational studies on members of the α/β hydrolase-fold family have shown that within the conserved core structure there is a catalytic triad of the form acid-histidine-nucleophile, and these amino acids have been shown to reside in similar topological locations within the different family members [143,144].

The key evidence suggesting the importance of Ser-135 as the key nucleophile in MhpC is the finding that it lies in a motif termed the "nucleophilic elbow" which is a short conserved sequence found in all the α/β hydrolase-fold enzymes which is of the form Sm-X-Nu-X-Sm-Sm, where Sm is a small amino acid (typically glycine), X is any amino acid and Nu is the nucleophile (see figure 9) [143,144].

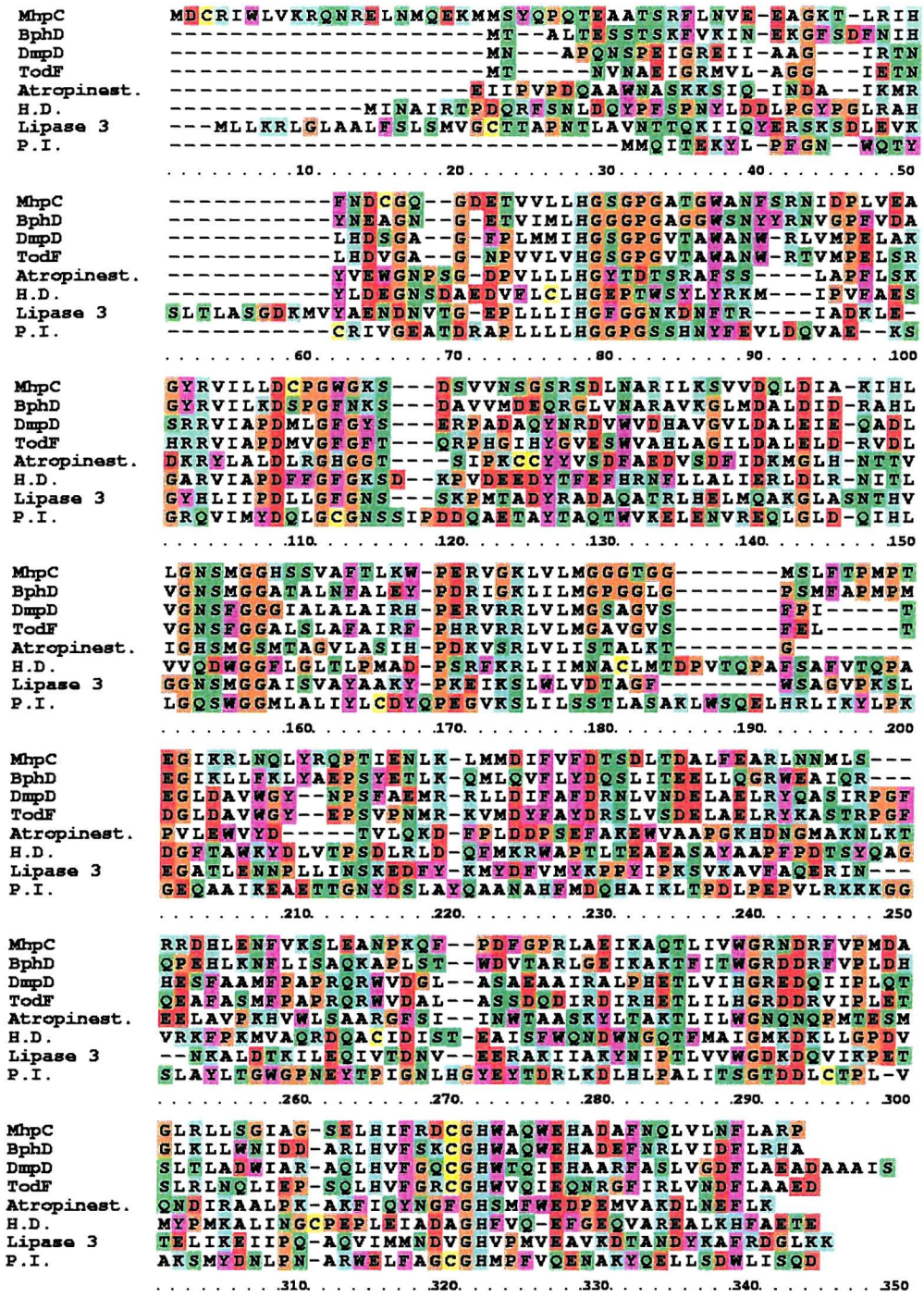


Figure 9. Primary sequence alignment of MhpC with several other hydrolase enzymes. Note the putative catalytic triad of serine-aspartate-histidine (positions 154, 293 and 322 in the alignment). These residues are precisely conserved in almost all of the sequences. Note that the serine lies in a conserved sequence ("the nucleophilic elbow"): Sm-X-Nu-X-Sm-Sm (see main text for details). Finally, note the degree of diversity between the different sequences, which indicates that there is only a conserved core of residues between the different enzymes. These features listed above are indicative of a group of hydrolases who adopt the so-called α/β hydrolase-fold. Sequences used: **1**. *E. coli* MhpC (2-hydroxy-6-ketonona-2,4-dienedioate hydrolase) [145,146,147,148]; **2**. *Pseudomonas cepacia* LB400 BphD (2-hydroxy-6-keto-6-phenylhexa-2,4-dienoate hydrolase) [149]; **3**. *Pseudomonas* CF600 DmpD (2-hydroxymuconic semialdehyde hydrolase) [150]; **4**. *P. putida* F1 TodF (2-hydroxy-6-keto-2,4-heptadienoate hydrolase) [151]; **5**. *P. putida* atropinesterase [152]; **6**. *Xanthobacter autotrophicus* GJ10 Halo (haloalkane dehalogenase) [153]; **7**. *Moraxella sp*. Lipase 3 [154]; **8**. *Lactobacillus delbrueckii* sub sp. *bulgaricus* proline iminopeptidase [155]. Sequences were obtained from the SWISS-PROT/TrEMBL protein sequence database [25]. The alignment was performed using MALIGN [20], and the colour presentation was generated by CINEMA (version 2.1) [18]. For a key correlating colours with residue types, see figure 1.

Ultimately, the importance of the putative catalytic triad in MhpC will be better understood once the hydrolase structure has been elucidated.

1.6 Aims of the work with MhpC

The primary aim of this work is to elucidate the mechanism of action of *E. coli* MhpC. To this end, our principal aim has been to solve the structure of the enzyme by X-ray crystallography, and to subsequently investigate the mechanism in more detail through inhibitor binding studies. Site-directed mutants will also be made since the enzyme may have the potential to bring about novel biotransformations, but the overall key aim of this work is to contribute to the understanding of the biodegradation of environmental pollutants.

Chapter 2

X-Ray Crystallography

Data Collection and Processing

2. X-ray Crystallography, Data Collection and Processing

2.1 X-ray Sources for Protein Crystallography

X-rays are waves of light (electromagnetic radiation) which are defined as having wavelengths in the range of 0.1-100Å [156]. As a consequence of their size, X-rays can physically interact with the electron clouds of atoms which are of similar dimensions to the wavelength. Atoms typically have diameters in the region of 1.5Å. From this interaction, structural information can be obtained detailing the three-dimensional positions of electrons, and hence atoms, in space.

X-rays may be generated in one of two ways:

(1) Electron Bombardment

When a high energy electron collides with a metal atom, an electron from an inner shell within the target atom may be thrust out, away from the nucleus. When this occurs, an electron from a higher orbital can then drop down to the lower level to displace the ejected particle, and as this happens, energy is emitted from the electron in the form of an X-ray photon [156].

(2) Electron Acceleration/Deceleration

When electrons (or positrons) travelling at certain velocities near the velocity of light are exposed to an accelerating (or decelerating) force, these particles will emit X-ray photons [157].

For the purposes of protein X-ray crystallography, two types of X-ray source are predominantly used. These are the rotating anode tube and the synchrotron (particle accelerator) source. Both of these are described below:

2.1.1 The Rotating Anode Tube

With this device, X-rays are primarily produced via electron bombardment, although a small proportion of X-rays will also arise from electron deceleration as the electrons collide with their target atoms. The rotating anode tube apparatus is summarised diagrammatically in figure 10.

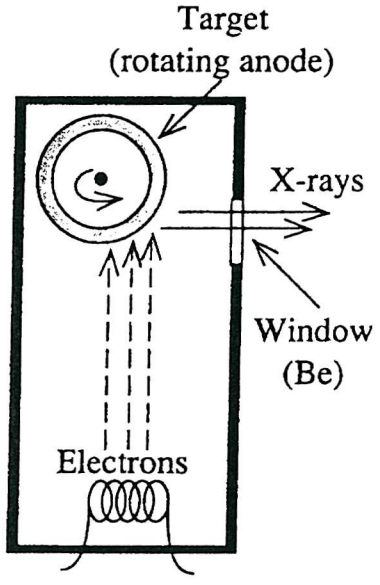


Figure 10. The rotating anode tube. The chamber illustrated here is evacuated when in use. Be is beryllium, an element which has low X-ray absorption and can thus provide a window through which the X-rays may pass on exit from the source. A coolant, (usually water) flows through the anode. Picture taken from "Crystallography Made Crystal Clear", Gale Rhodes, 1993 [156].

Initially, a filament is heated which serves as an electron source. Electrons are then accelerated across a potential difference (typically around 40kV, along with a current of 37mA for a 1.5kW tube) which is set up between the filament (cathode) and a metal cylinder which is usually made of copper (anode)[158]. Upon impact with the anode, the electron

beam then elicits release of X-rays, principally through electron transitions within the target metal as discussed above. The intensity of the resultant X-ray beam is, however, limited by heat produced during the electron bombardment at the anode, although a coolant such as water is used to help dissipate the thermal energy away. This heating of the anode restricts the power load which can be applied to the electrons in order to make them accelerate from the filament. By rotating the anode, however, the heat poses less of a problem as the anode metal is further cooled in-between exposures to the electron beam. Thus a higher power may be applied to the electrons passing between the electrodes and an X-ray beam of greater intensity can be obtained [158].

For a copper anode, when electrons from the K-shell are boosted out of this orbital, then electrons from the L- and M-shells move into the K-shell to fill up the vacancies there, and in doing so, release energy in the form of the X-rays. Electrons passing from L- to K-shell emit so-called K_α radiation. For copper, this is released at two distinct wavelengths: $K_{\alpha 1}$ = 1.54051Å and $K_{\alpha 2}$ = 1.54433Å. Two discernible wavelengths arise here due to L-shell electrons existing in two separate energy levels within the L-orbital, which means that radiation of two different energies, and thus wavelengths, will be released upon transition to the K-shell [158].

However, despite the production of these two X-ray wavelengths, copper K_α radiation generated by a rotating anode tube is assigned a single wavelength by convention since the effective wavelength of the overall K_α radiation is the weighted average of the $K_{\alpha 1}$ and $K_{\alpha 2}$ values, which equals 1.54178Å, since the intensity of $K_{\alpha 1}$ is twice that of $K_{\alpha 2}$ [158].

Electrons in the anode passing from M- to K-shell give rise to K_β radiation of single wavelength. For copper, this equals 1.39217Å. Only one wavelength is effectively emitted from such transitions due to the M-shell energy levels being very close together [158].

A summary of these electron transitions within the anode upon excitation due to electron bombardment is shown in figure 11.

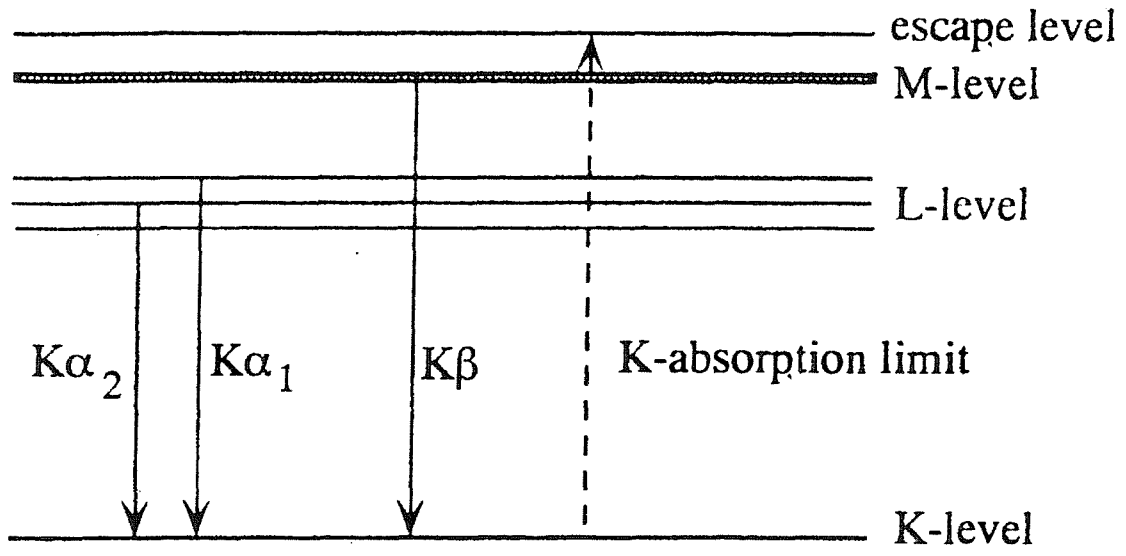


Figure 11. Diagram summarising the electron transitions that generate X-rays from a metal upon excitation via electron bombardment. The K-, L- and M-levels represent electronic orbitals within a copper atom. Taken from "Principles of Protein X-Ray Crystallography", Jan Drenth, 1994 [158].

An important feature of an X-ray source as far as the protein X-ray crystallographer is concerned is that X-rays are output at a single wavelength. This is because X-rays at different wavelengths will interact with a target crystal differently, thus giving rise to unique diffraction patterns corresponding to each incident wavelength used which may partially overlap one another, thus rendering the diffracted spot intensities unmeasurable.

The X-ray emission spectrum of an X-ray tube with a copper anode is shown in figure 12. From this it can be seen that X-rays are not only produced as the aforementioned K_α and K_β radiation, but are also released at other wavelengths as well. This additional background radiation (shown by the broad peak centred around about 0.7Å in figure 12) is called "Bremsstrahlung", and is the result of electron deceleration (or acceleration) events occurring within the X-ray tube [158].

In order to select a single wavelength from this emission spectrum, one of two devices are used: either (1) a filter, or (2) a monochromator crystal.

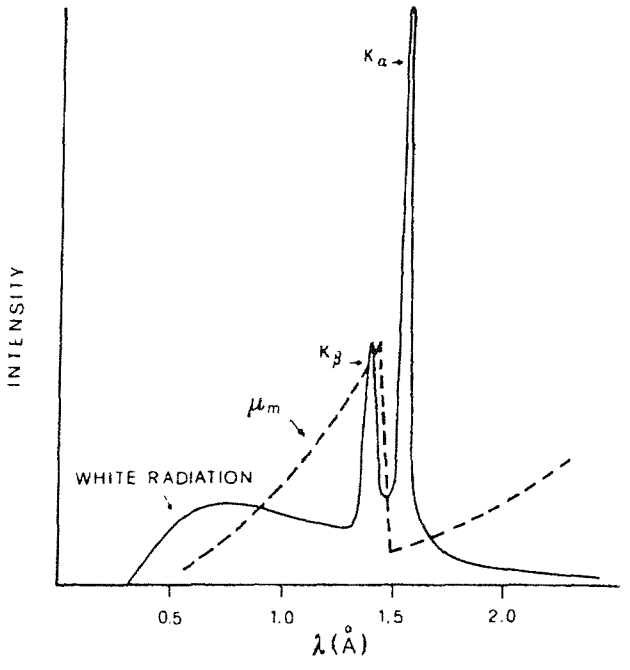


Figure 12. X-ray emission spectrum of an X-ray tube with a copper anode (continuous line). The distinctive peaks derive from electron transitions within the copper atoms, whilst the continuous "Bremsstrahlung" region (labelled "white radiation") arises from electron deceleration (or acceleration) events. The relative intensity of the X-rays emitted at each wavelength is shown on an arbitrary scale. The dashed line labelled "μm" represents the absorption spectrum of nickel. Notice how nickel can be used to filter away the copper K_β peak, whilst leaving the K_α emittance strong. Picture taken from "Fundamentals of Crystallography", Monaco, 1992 [157].

For X-rays derived from a copper anode, a nickel filter may be used. Nickel has an absorption edge at a wavelength just above that of the copper K_β radiation (see figure 12). This absorption edge corresponds to a dramatic dip in nickel's X-ray absorption spectrum which is flanked at low wavelength by a region of steadily-increasing absorption up to the edge, and at high wavelength, after a small lag period, by an area of further-increasing absorption. The positioning of this peak and trough of the absorption edge for nickel are so-placed as to maximise the absorption of K_β radiation output by the anode, whilst minimising the absorption of K_α radiation. Thus, the nickel can be used as a filter, effectively removing the copper K_β radiation and also part of the Bremsstrahlung region as well. The end product of

this arrangement is the production of monochromatised copper K_α radiation. (As a general rule for elements with an atomic number ≤ 70, if an element of atomic number Z, emits K_α and K_β radiation, then the element of atomic number Z-1 will have an absorption edge around a wavelength that allows strong absorption of the K_β radiation, whilst minimising absorption of the K_α radiation) [157]. Although highly monochromatised, the copper K_α radiation released by such a filter still has a residual background spread of radiation of different wavelengths associated with it. In order to improve this situation further, a monochromator crystal is sometimes used as an alternative to the filter. The monochromator used for the rotating anode tube is a graphite crystal [158]. The principle behind its function lies in Braggs' Law. This tells us that when X-rays of different wavelength are diffracted by a crystal, the diffraction angles of the diffracted beams are dependent on the individual wavelengths of each of the beams. Thus, a particular X-ray beam of desired wavelength can be selected by simply choosing a particular diffraction angle.

Crystal monochromators produce much cleaner single wavelength X-ray radiation than filters, and in some instances have been able to resolve even the $K_{\alpha 1}$ and $K_{\alpha 2}$ doublet which are inseparable by any filtering method [157].

After their genesis, the X-rays leave the rotating anode tube through X-ray transparent beryllium windows [158]. These windows are needed because the entire rotating anode chamber is evacuated, and hence a sealed environment. A vacuum is necessary to prevent collisions between the accelerating electrons emerging from the filament and air particles, which would significantly weaken the efficiency of the X-ray generation process [157].

Beyond the X-ray source, the beam then journeys through a series of devices which neaten it up in preparation for its rendezvous with the crystal. Firstly, the beam passes through the

aforementioned filter or monochromator, it then travels on to a collimator and sometimes passes over a mirror en route to the target crystal.

The collimator is a simple device used to keep the X-ray beam narrow and the individual X-rays as parallel as possible. It is normally a hollow metal cylinder which has two apertures defining the beam size, and a third that removes scattered radiation originating from the defining aperture most distant from the X-ray source [157]. Whatever the source of X-rays, a collimator is always required [156].

Finally, after collimation, focusing mirrors may be employed to reduce the beam diameter still further [156], and the end product is a monochromatised, focused beam of copper K_α radiation which then heads off towards the crystal.

2.1.2 The Synchrotron

To the protein X-ray crystallographer, the synchrotron, as a source of X-rays is the optimal choice. This is primarily because the X-rays produced are more concentrated together which provides an X-ray beam of high intensity (typically at least two orders of magnitude greater than that of a conventional X-ray tube [158]). This allows rapid data collection, e.g. data taking several days to acquire on a conventional X-ray source, may be obtained in just 2-3 hours, and has the added benefit that radiation damage to crystals may be reduced due to the far shorter exposure times. This is particularly useful when collecting Multiwavelength Anomalous Diffraction (M.A.D.) data, where a single crystal is ideally used for the collection of multiple data sets in order to minimise scaling errors and anisomorphic effects, i.e. inherent differences between one crystal and the next.

The availability of a highly intense X-ray source provides additional benefits to the protein – more so than to the small molecule – X-ray crystallographer, since proteins can be very large, and consequently can have relatively few diffracting particles within a given crystal. Since the

strength of diffraction is proportional to the number of diffracting particles, then from crystals comprising large proteins, the diffraction is relatively weak. A more intense X-ray source can provide stronger diffraction from such crystals per unit time. This theory can also be applied to small crystals having, once again, a relatively small number of particles inside. Moreover, the greater intensity of the X-rays is also beneficial to the X-ray protein crystallographer since proteins are composed of light elements such as hydrogen, carbon, nitrogen and oxygen, which possess few electrons per atom and as such, are relatively poor at diffracting the X-rays [158].

In addition to the advantages of using synchrotron radiation in protein X-ray crystallography associated with the beam intensity, extra benefits arise due to the wide spectrum of strong X-rays which are generated from a synchrotron source. Firstly, this means that X-rays of a high energy and low wavelength may be used for data collection which are not available with conventional X-ray tubes. At these wavelengths, X-rays can probe target crystals for higher resolution structural information by sampling them at finer and finer intervals. This has a profound impact on the atomic detail achievable from a given crystal. Another advantage of the availability of low wavelength X-rays at synchrotron sources is that absorption of X-rays by atoms (below absorption edge wavelengths) decreases as the wavelength shortens. This absorption of X-rays by crystals is damaging due to free radical formation, which can subsequently spread throughout the crystal, thus disrupting the ordered structure. It is therefore in the best interests of the X-ray crystallographer to keep absorption effects to a minimum [158,157].

A final key advantage of synchrotron sources is that at many of them the experimenter is able to select desired wavelengths. This tuneability is of particular importance where M.A.D. experiments are performed, since data sets must be collected at several unique wavelengths using this method [158].

The basic apparatus of a typical synchrotron radiation source is summarised in figure 13. X-rays may be derived from either electrons or positrons using this type of source, where positrons offer the attraction of longer beam lifetimes and greater stability, but for simplicity and financial reasons, electrons are most commonly used [159]

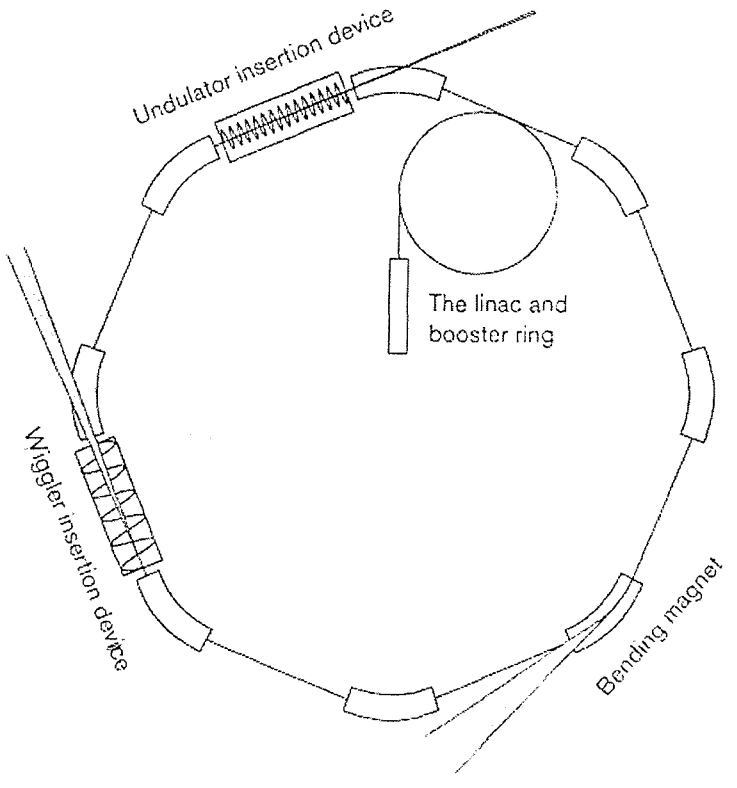


Figure 13. Anatomy of a synchrotron source. Examples range from ten to a few hundred metres in diameter. Electrons/positrons are accelerated using a linear accelerator (LINAC) and sometimes by an additional booster ring as well, and are then transferred to a storage ring. This contains several different magnetic devices which bring about particle beam acceleration and thus X-ray generation. Picture taken from reference [159].

The initial stage in the overall process is the generation and acceleration of the electrons/positrons up to velocities approaching the velocity of light. This is important since the velocity of the charged particle determines the energy of the electromagnetic radiation which is emitted from it upon exposure to an accelerating force. Consequently, only when velocities nearing that of light are attained, can such particles emit electromagnetic radiation having energies in the X-ray region of the overall spectrum [157].

Particle acceleration is achieved using a linear accelerator (or LINAC). Electrons are generated from such a device and are then accelerated by electric fields. Positrons, where used as an alternative, are produced when such an electron stream impacts upon a tungsten target [159].

A LINAC is sometimes used to accelerate these particles to their ideal velocities for X-ray emission, but, more often, a secondary system called a booster ring is employed to further-accelerate the particles output from the LINAC up to their desired velocities. The final energies of the accelerated particles should be in the region of 1-8 GeV [159]. It is at this stage that the particles enter the all-important storage ring which is where the X-rays are released for experimentation. The storage ring is, in reality, a polygon with corners and straight sections. The corners house bending magnets which guide the electron/positron beam around the central cavity, whilst the straight sections contain insertion devices [159]. These will be briefly discussed shortly. At certain positions around the storage ring, beam lines branch off at tangents to it, and around each of these beam lines are built work stations where experimental data may be collected from crystals.

X-rays emerge from the storage ring when the electrons/positrons continually circulating within the ring cavity experience acceleration. This happens when the charged particles are forced to deviate from their circular path by magnets. This shift in direction is accompanied by a centripetal acceleration (that is, towards the centre of the ring) and it is this acceleration that brings about the X-ray emission. After the release of X-rays, the accelerated particles return to their original path, and they then continue their journey around the ring along with all the other particles as before [158,159]. A consequence of the X-ray emission by the charged particles is that energy is lost from the electrons/positrons during the X-ray release. This energy must be restored to the particles in order to ensure they maintain their near velocity of light passage around the ring. To this end, an energy booster is present in the

synchrotron ring which takes the form of a radio frequency alternating electric field input that is applied to the particles every cycle [158]. In order to prevent these fluctuations in velocity that particles experience as they move around the storage ring from disrupting the circular particle beam path, there must be a synchronisation between the changing velocity of the circulating particles and the relative strength of the bending electromagnets placed at the corners of the polygon which control the particle beam position. It is from this synchronisation process that the synchrotron source derives its name [160].

Throughout the synchrotron, wherever electrons (or positrons) are present, a vacuum exists. This is necessary to prevent collisions between electrons/positrons and air particles once again, which would result in the loss of electrons/positrons and thus a disruption of the particle flow [157]. Despite the presence of the vacuum however, residual air particles will always remain and accordingly, as time passes, electrons/positrons will interact with these and so will be lost. Therefore, a storage ring containing charged particles will gradually lose the beam current over a time period until a point is reached where the current flow is so low that the beam must be dumped and the storage ring refilled with high velocity particles. Beam dumping/refilling can happen several times a day or at very best once a week. During beam dumping and refilling, clearly the synchrotron will not produce X-rays, so experiments are put on hold during this period [159].

Which parts of the synchrotron yield usable X-rays? As discussed above, all sections of the storage ring where particles are forced by magnets to bend away from the direction of the main beam and back again represent the regions of X-ray release. Thus, X-rays can originate from particle interaction with the bending magnets located at the corners of the storage ring polygon. Such X-rays are emitted in a wide spectrum of energies and accordingly may be tuned to a particular wavelength by an operator using them [159]. These X-rays, despite being intense, are less so than those produced elsewhere around the ring, in the insertion devices

[158]. The insertion devices are sections of the storage ring where a series of several magnets of alternating polarity are found. They are divided into two categories: the wigglers and the undulators.

In a wiggler, the series of electromagnets of alternating polarity are relatively strong and few in number. Electrons/positrons passing through such a device maintain their forward trajectory, but simultaneously oscillate sharply from side to side about the theoretical main particle beam centre, depending on the polarity of each of the magnets they pass. The peaks of the X-ray release correspond to the points where the particles reach their maximal amplitudes during their journey through the wiggler [158]. The radiation released from each peak or trough along the particle's path serves as an individual X-ray beam which is totally independent from all others produced in the wiggler, i.e. there is no constructive interference between the different X-ray beams produced in such a device [157]. The X-ray beams output from a wiggler collectively emerge as a wide spread of radiation which horizontally fans out, and which travels in the general direction of propagation of the electrons/positrons. Several work stations can be fed from such a beam, each taking a slice of the radiation [159]. A wide spectrum of X-ray radiation is output by a wiggler, and accordingly such a device produces wavelength-tuneable X-rays [159].

In contrast to the wiggler, an undulator has more moderate magnetic fields guiding the particle beam through it, and as such, the curvature of the particles' side-to-side path through the undulator is far less sharp [158]. As a consequence, X-ray beams produced from such a device are released in a very similar direction to one another, once again, heading in the general direction of propagation of the main electron/positron beam, but this time they physically interact with one another. Thus, radiation output from an undulator experiences interference effects. This leads to elimination of many of the X-ray wavelengths originally output at each turn of the electron/positron beam due to destructive interference, whilst X-

rays at certain key wavelengths summate through constructive interference and thus emerge from the interacting individual X-ray beams as distinct X-ray wavelengths within the output beam [159]. One of these wavelengths is far stronger than the others due to its experiencing particularly favourable interference effects, and hence the output X-ray beam from an undulator is effectively monochromatised. Thus, an undulator is not a wavelength-tuneable X-ray source [158].

In contrast to a wiggler, the extremely narrow X-ray beam output by an undulator is too focused into a single beam to provide X-rays for a series of work stations arranged in a fan around the source. However, multiple work stations may be fed from such an X-ray beam using X-ray transparent crystals (such as diamonds) as monochromators to produce secondary beams [159]. A positive consequence of the undulator X-ray beam divergence being so low is that diffraction spots produced from a crystal exposed to such radiation will generally be extremely well resolved, i.e. sharp spots with minimal smearing [158].

The various properties of a wiggler and an undulator are compared in table 3 below.

	WIGGLER	UNDULATOR
Quantity of bending magnets present in device	Relatively few	Many
Magnetic Field Strength	High	Moderate
Sharpness of bends in the electron trajectory as these particles pass through the device	Very sharp	Moderate
Presence of interference effects	No	Yes
Appearance of emerging beams	Wide beam of radiation	Highly focused beam arising from many individual beams
Wavelengths output	Wide spectrum of X-rays produced in each beam	One major X-ray wavelength produced along with a few distinct (minor) others
Intensity of emitted radiation	High	Very high
Tuneability of emitted radiation	Wavelength-tuneable	Not wavelength-tuneable

Table 3. A summary of the differences between a wiggler and an undulator

Once the X-rays have been generated, they then normally need to be monochromatised, passed through a collimator and then over a focusing mirror as described above for the rotating anode tube.

Monochromators for a synchrotron source are normally crystals of germanium or silicon, since these select out a very narrow X-ray wavelength band from an incident beam [159]. Such monochromators may be of either the single or double variety. The former can be flat or bent, the bent type being advantageous since they can help with focusing of the X-rays. The single type monochromators have a disadvantage as far as wavelength-tuneable synchrotron

sources are concerned. Tuning such monochromators to a new wavelength alters the scanning angle of the monochromator which means that the whole of the X-ray diffraction equipment must be moved to accommodate this change. To overcome this problem, a double monochromator is used, where two monochromator crystals can act together to diffract the beam twice so that as the wavelength is changed, the net beam diffracted by both of these crystals is always output in the same direction [158].

It should be pointed out that some experiments performed at the synchrotron such as Laue work, require all of the wavelengths generated by the source, and as such, these experiments do not need a monochromator [159].

Collimation at a synchrotron is achieved using slits to eliminate unwanted scattered radiation, and mirrors made from, for example, rhodium-coated silicon or glass, are also used to focus the X-rays onto their target crystal [159].

2.2 Detectors for X-ray protein crystallography

When an X-ray beam strikes a target crystal, it becomes split up by diffraction into many beams which emerge from the crystal as continuous rays of energy that fan out away from one another as they travel through the air. Each diffracted X-ray beam contains structural information derived from all electrons (and thus atoms) found within the crystal, and as such, the X-ray crystallographer needs to be able to capture as many of these beams as possible and measure their relative intensities in order to gain access to the information which they contain. To this end, a variety of detector devices have been designed which employ many different methods of X-ray detection. Despite this diversity, however, all of the detectors strive to satisfy certain key criteria with respect to collecting the X-ray information.

Sensitivity, is such an important criterion, and detection of individual X-ray photons is the ultimate goal here. Of equal importance is dynamic range, which is the intensity range of diffraction spots collected on a detector surface which may be accurately recorded in a single diffraction pattern. Ideally this intensity range should be as wide as possible so that both weak and strong diffracted beams can be recorded on a single diffraction image, thus maximising the data collected.

Accuracy of detector response to X-rays is another key factor, i.e. it is important to ensure that a uniform signal is both registered and subsequently output by a detector in response to the X-rays that strike it.

Speed of signal output is yet another consideration. If a detector can transfer a collected image into stored computer data rapidly then this has implications in particular for synchrotron trips where only a limited time slot is available, and thus, the faster the detector fulfils its role, the more data can be collected on such trips.

Finally, the resolution of the detection process is also important so that individual diffraction spots have every chance of being identified as separate entities.

These, although not an exhaustive list, are some of the more important features that a detector should exhibit.

Despite the many designs which exist for X-ray detectors, in this body of work, only two types of detector have been used: the image plate and the charge-coupled device, and as such, only these two detectors shall be discussed below.

2.2.1 The Image Plate

Using this detector, the X-ray energy output from the numerous diffracted beams is first stored within a layer of photostimulable phosphor crystals, and it is then released as visible light and converted into digital data using an image scanning device. The details of the overall process are outlined below:

The phosphor crystals which lie at the heart of this detection process are composed of a barium halide phosphor (BaFBr) and also europium ions (Eu^{2+})[161]. These crystals are mounted in an approximately 150μm thick layer onto a flexible plastic base using an organic binder [162]. Upon exposure to X-rays, europium electrons within these crystals become excited to higher energy levels, forming Eu^{3+} ions in the process [158,162]. As this happens, a significant proportion of the X-ray energy imparted to the crystals becomes stored in them, as europium electrons become trapped in so-called colour centres (or F-centres) [158,162]. Energy stored in this way is very slowly released over a number of days (the energy decays with a half-life of about eight hours at room temperature)[158,162], thus a diffraction pattern representing the relative intensity of each of the different X-ray beams output from the protein crystal is stored in the detector phosphor crystals at this stage in a semi-stable form.

Next, the diffraction pattern must be scanned pixel by pixel and the information converted to digital data. The stored energy may be released from the photosensitive phosphor crystals by exposure to a visible light source. A red helium-neon laser beam is used for this, having a wavelength of approximately 633nm, and results in the emission of violet luminescence of about 390nm wavelength [161]. This violet light comes from the electrons trapped in the colour centres, and emerges as they are returned to their europium ions [162].

The red laser light is filtered away during scanning of the image plate, whilst the violet light is collected using a photomultiplier tube [158]. The emitted light is of a wavelength where it may be detected by the photomultiplier tube with high quantum efficiency [162], and thus, with such a wavelength, signal detection during scanning of the phosphor image plate, may be achieved with high sensitivity.

The photomultiplier used in scanning is mounted within a moveable reading head [159], and scans over the surface of the phosphor image plate by moving from the outside to the centre, covering a radial distance, whilst the plate itself spins around in a particular direction. As such, the diffraction image is read from the plate as a spiral of data. The electronics can then convert the measured fluorescence emission from each part of the scanned image into digital information which can thence be analysed using specialised computer software.

Once the image plate has been scanned, it must finally be erased before data collection may proceed with the next image. Erasing the plate is necessary because residual electrons remain trapped in the colour centres after scanning, and as such, they need to be returned to the Eu^{3+} ions before collecting the next image so that no background contamination is carried over from one diffraction pattern to the next [162]. This erasing process is very simple, involving exposure of the image plate to bright visible (optimally yellow) light [161], output by two erasing bulbs, and once completed, data collection may then proceed.

The image plate detector has several advantageous features. Principally, it is more than ten times as sensitive as conventional X-ray film, and it has a far superior dynamic range in the region of $1:10^4$-10^5 (i.e. it can detect a range of spots in one image where both the very weakest spot, with an arbitrarily assigned intensity value of one, and the very strongest, having a relative intensity of 10^4-10^5 times greater, can be collected and measured accurately) [158]. Another benefit with the image plate is that background noise is very low indeed. This is partly a consequence of the phosphor plate (and hence background scatter) being erased moments before each image is collected. The noise level can be so low in fact that it may represent less than three X-ray photons per $100\mu m^2$ detector area [163].

The efficiency of the image plate is another very important feature. This is a measure of how many of the X-ray photons incident upon the detector surface (minus background scatter) are converted to output digital data. For an image plate, this efficiency is extremely good, being around 70%, a figure which holds true across the full dynamic range and which is kept at this level over a wide spread of incident X-ray wavelengths [161].

In addition to all of these advantages, the image plate has a few noteworthy disadvantages. The most important of these is its speed. This is because the scanning/erasing processes are very time consuming, together taking between 40-150 seconds per image (depending on detector type, image and pixel size)[159]. Thus over the full course of data collection, where typically around a hundred images will be acquired, a considerable proportion of time is spent scanning/erasing the image plate which slows the overall process quite dramatically.

Another potential drawback with the image plate is that the diffraction pattern stored in the phosphor crystals prior to scanning is only semi-stable. In fact, the stored image begins to fade away quite rapidly over the first few minutes after X-ray photons are detected, but this process then slows down significantly [158]. In practice, this fading of a stored image is rarely problematic [163].

A final potential disadvantage, worth only briefly mentioning, is that the spatial resolution of the image plate is not ideal, being around 100-150μm. X-ray film, by comparison, provides an effective resolution of around 20μm. However, once again, in most cases this does not present a problem to data collection [161].

In the overall balance of things, the image plate is a very effective device with only minor disadvantages. As such, it is one of the most commonly used detectors involved in crystallographic research today.

2.2.2 The Charge-Coupled Device (CCD) Detector

The charge-coupled device (CCD) detector is the most advanced detector-type used in X-ray crystallography today. There are several different versions of such a detector available, but each of these has a CCD sensor at its heart [164].

There are two basic design templates for these devices. Firstly, there is an arrangement where X-rays strike the CCD directly. This setup, however, has certain problems. The CCD is susceptible to radiation damage; signal saturation can readily occur as a consequence of each X-ray photon depositing high energy on the CCD and the efficiency of such a device is low [164]. As a consequence, this design is not often seen.

The second design template incorporates an X-ray-sensitive phosphor surface at the front end of the CCD which shields the latter from the radiation, and which converts the X-ray signal to visible light immediately (in contrast to the stored energy used in the image plate) after radiation exposure. This immediate signal transfer from X-ray to visible light ensures the overall signal detection process is very rapid indeed. An example of a phosphor used is Gd_2O_2S:Tb (also known as P43), which is one of the most efficient known, converting around 15% of its incident X-ray energy into visible (green) fluorescence. Its speed is quite rapid as well: After about 10 milliseconds, light emission drops to less than 0.1% of its initial value

[164]. The visible light emitted by such a phosphor is passed along a fibre-optic taper (comprised of many tiny glass tubes with very narrow bores which taper at their ends most distant from the phosphor plate) which serves to maximise the signal transfer from phosphor plate to CCD whilst demagnifying the data (i.e. taking the information emitted by the large phosphor plate and reducing it to such a degree that the smaller CCD can respond to all of it) [164]. An example of such a setup is shown in figure 14.

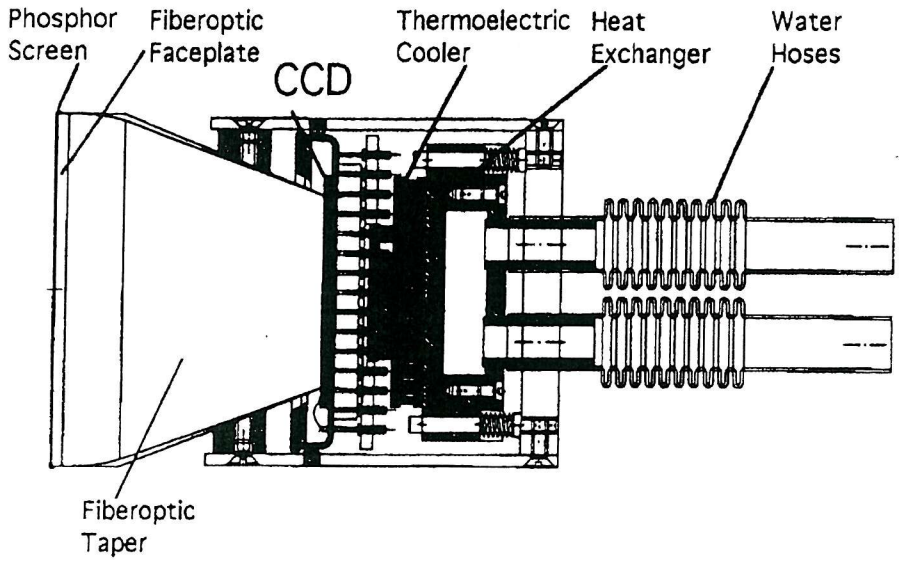


Figure 14. Anatomy of a typical CCD detector. This is fronted by a phosphor plate-fibre-optic taper which collects and conducts the structural data to the CCD chip where it is converted to a digital signal. The CCD is cooled to around -40°C to dramatically reduce the spontaneous generation rate of electrons which appear due to thermal factors, and which contribute to background noise within the CCD. The heat is transferred away by water [164]. Picture taken from reference [164].

The CCD itself functions in two phases: firstly, it acquires and accumulates charge in response to incident X-rays or visible light derived from a phosphor, and secondly, after a period, it then relocates this charge, quantifies it, and converts it into digital signal [164]. The device is composed of a silicon-based semi-conductor material and is essentially arranged into two distinct sections: firstly there is the so-called vertical (or parallel) register,

which is a two-dimensional spread of photosensitive pixels, each of which has the capacity to store charge, and secondly there is the horizontal (or serial) register, which is a separate row of pixels where accumulated charges can be shifted one by one into an analogue-to-digital converter [162].

The basic CCD arrangement is summarised in figure 15.

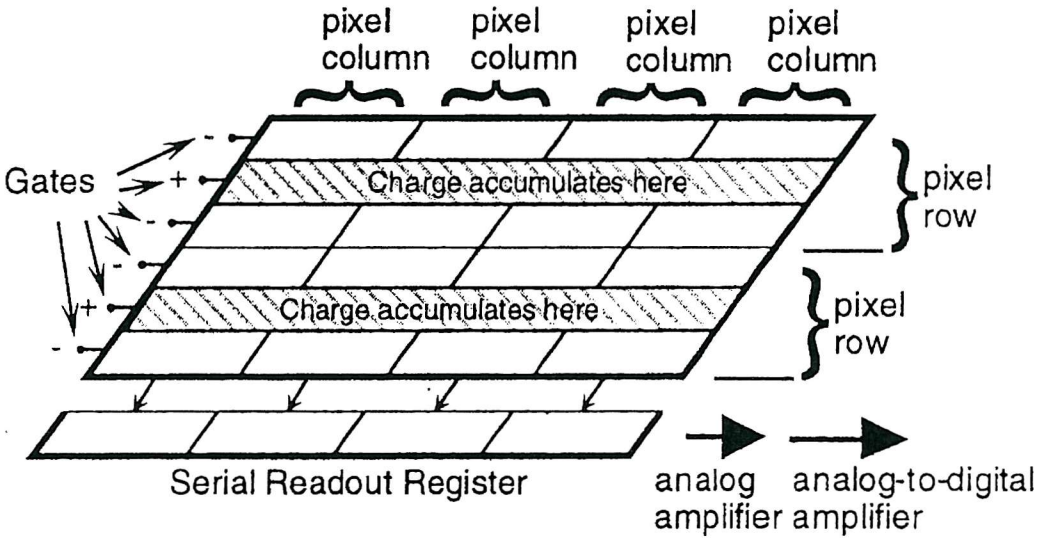


Figure 15. Diagrammatic representation of a CCD. The diffraction spots are initially stored as clusters of charge in the two-dimensional pixel array termed the vertical register. Measurement of these accumulated charges is subsequently achieved by passing them to the horizontal (serial readout) register which then feeds them through an analogue-to-digital converter for quantitation. Taken from reference [164].

The fine structure of the CCD is made up of a series of electrodes termed "gates". These are arranged into parallel rows (as shown in figure 15) and are organised so that every third gate is biased at a positive potential, whilst the remaining two are biased negatively. Collectively, three adjacent gates comprise a pixel, and when photons of light strike a pixel, electron-hole pairs are generated by the photoelectric effect. Electrons released in this way then migrate to the positive electrode (gate) within the pixel closest by, where they are stored. The gates set at negative potential either side of the positive gates prevent the accumulated electrons in a

given pixel from leaking away to an adjacent row, whilst pixel columns are charge-isolated from one another by thin channel-stop strips which have high boron content. Thus, each pixel's charge is effectively segregated from that of its surrounding neighbours [164]. This, then, is how charge is acquired and stored in the CCD in response to light.

Once an entire diffraction pattern has been captured, data collection then stops as the CCD converts all of the stored charge accumulated in the pixels into digital information. Initially, all of the rows of the vertical register containing the pixels with the diffraction data move their stored charges from one gate to the next towards the horizontal register. During this process, the last entire pixel row from the vertical register shifts its charges (still as distinct entities) onto the horizontal register, and then the pixel charges in this row alone are then again shifted, this time sideways, towards an electronic circuit. Upon contact with the electronic circuit, a pixel charge is measured for conversion to digital data, and once completed, an adjacent pixel charge in the same row is then passed along for quantitation. This continues until all charges on the horizontal register have been measured, at which point, the vertical register once again shifts the pixel charges by one row towards the horizontal register, and the overall process repeats until all pixel charges in the vertical register have been recorded. Charge shifting within the CCD is achieved via a simultaneous lowering of the positive electrode voltage at the centre of each pixel where the charge is stored to a negative value, whilst increasing the voltage of a neighbouring electrode to a positive value. The electrons corresponding to the charge in each pixel will thus move towards the newly formed positive electrode and will become stored there instead. The efficiency of this transfer process in modern CCDs is greater than 99.9999%, and thus any losses during transfer are not all that significant [164].

As each pixel charge is passed to the electronic circuit, it is first dealt with by an on-chip, very low noise preamplifier, then secondly by an amplifying analogue processing circuit and

finally it is stored as digital data by an analogue-to-digital converter [164]. Thus, the X-ray data can be accumulated in this way for subsequent processing on a computer.

The main advantage of the CCD-based detectors is that of speed. Each image can be typically read into digital signal by such a device over just 1-10 seconds. This has had a dramatic impact at synchrotron sources where available beam time for research groups is limited [159].

The CCD-based detectors also have extremely low background noise associated with them. These levels are so low in fact that each individual X-ray photon alone generates a signal in the detector which equals or is greater than the total detector noise. This allows such devices to near the point at which they can detect single X-ray photons [164].

Another important feature of the CCD-based devices is their dynamic range which may be around $1:10^5$, and thus comparable to the image plate [162]. Such a wide dynamic range is advantageous for all spot intensity measurements, but of particular note is the effect that broad range can have on multiwavelength anomalous diffraction (M.A.D.) phasing experiments, where small differences between certain measured intensities must be accurately recorded, i.e. the wider the dynamic range, the greater the accuracy of the measured intensity differences [164].

One other potential benefit of the CCD-based detectors is their resolution. This is around 40μm which approaches the spatial resolution of X-ray film, and thus is more powerful than the image plate in this regard [162]. However, how much additional benefit this high degree of resolution affords the CCD-based detectors is questionable.

The most noteworthy drawback with the CCD-based detectors lies in their relatively small physical size, and hence detection area [159]. This limits the resolving power of the detector since such a device must be placed close to the sample crystal if it is to capture the high resolution X-rays of wide diffraction angle, and as such, the spots collected are very close together. This problem can be circumvented by using several CCD-based detectors 'strapped'

together in either a 2x2 or 3x3 arrangement. The physical gap between detector subunits in such a grid can be reduced to typically 2-3 pixel widths, and therefore minimal data will be lost in these gaps where such a setup is used. Unfortunately, there is a cost limit to how many of these detector units can be added together in this way [164].

2.3 X-ray Diffraction Data Processing: Theory and Programs:

2.3.1 Data Processing: An Overview

The key role of the X-ray crystallographer is to interpret the X-ray diffraction patterns produced by a given crystal in the correct way, and using the best tools from those available, in order to accurately determine the three-dimensional positions of electrons, and thus atoms, within the crystallised protein. The overall process from diffraction pattern to completed structure may be broken down into three distinct stages:

(1) The determination of the **structure factor amplitudes** of all the diffracted X-ray beams which give rise to the dark spots seen on the X-ray diffraction patterns (i.e. a measure of the strength of each of these diffracted X-rays)

(2) The determination of the **phase** component of each of these diffracting X-ray beams (i.e. elucidating how the net X-ray waveform appears as it travels through space), and

(3) **Refining** the resultant electron density maps to provide the closest fit with the experimental data.

Equation 1 below is termed the electron density equation which summarises the fundamental task of the X-ray crystallographer.

$$\rho(xyz) = 1/V \sum_{hkl} |F(hkl)| \exp [-2\pi (hx + ky + lz) + i\alpha (hkl)] \qquad \text{Equation 1}$$

$\rho(xyz)$ represents the electron density at any particular point in the unit cell (the repeated building block of the crystal), V is the unit cell volume, $|F(hkl)|$ is the aforementioned structure factor amplitude term of reflection (hkl), including the temperature factor, which describes the amplitude of the net X-ray wave for a given diffraction spot (hkl) at the precise

point where the wave strikes the detector, x, y, z are the coordinates in the unit cell and $\alpha(hkl)$ is the phase (angle) of the net X-ray wave for each diffraction spot (hkl) as the wave impacts the detector. This equation serves to emphasise that in order to determine electron density at all coordinates in the crystal unit cell, the structure factor amplitude and phase information must first be obtained.

The **structure factor amplitude** for each of the diffraction spots is simply acquired from the spots themselves due to the following relationship: the structure factor amplitude is proportional to the square root of the integrated intensity of spot (hkl).

The **phase** term is not directly obtainable from the diffraction data however, and thus additional techniques must be employed in order to calculate this value. For example, multiple isomorphous replacement (M.I.R.) is a method historically popular where electron-dense heavy atom derivatives of the crystal concerned are prepared, whose diffraction patterns, when related to the native patterns, can provide the positions of the heavy atoms in the unit cell, from which all the phase information can thence be obtained.

A second very commonly used method for phase determination is molecular replacement. This relies on the existence of a protein homologous to that under study, whose structure has been solved. The search model can be superimposed in three-dimensional space onto the unknown structure whose phase information is, in effect, then applied to the protein under study whose structure can then be solved.

A third method of obtaining the phase information is via multiwavelength anomalous diffraction (MAD) phasing. Here the phase information is provided by electron-dense atoms such as selenium in the form of selenomethionine, which can be engineered into the protein of interest in the place of native methionine residues. The electron-dense selenium atoms within the modified protein are able to interact in special ways with the X-rays generated at certain distinct wavelengths so as to produce sufficiently different diffraction data at each

wavelength to allow their atomic positions within the unit cell to be pinpointed. Thus, the phases, as with the MIR method, can be obtained from this once the seleniums have been located.

The final step to structure determination is the **refinement** stage. Here, the electron density maps are modified via several cycles involving (i) modelling atoms into three-dimensional graphical representations of the electron clouds, then (ii), using refinement software to optimise all bond lengths and angles and thence compare how well the output model agrees with the experimental diffraction data. This process is repeated many times until the model fits the experimental data optimally.

A detailed examination of how data processing was performed with both momordin and MhpC follows.

2.3.2 Data Processing - Momordin

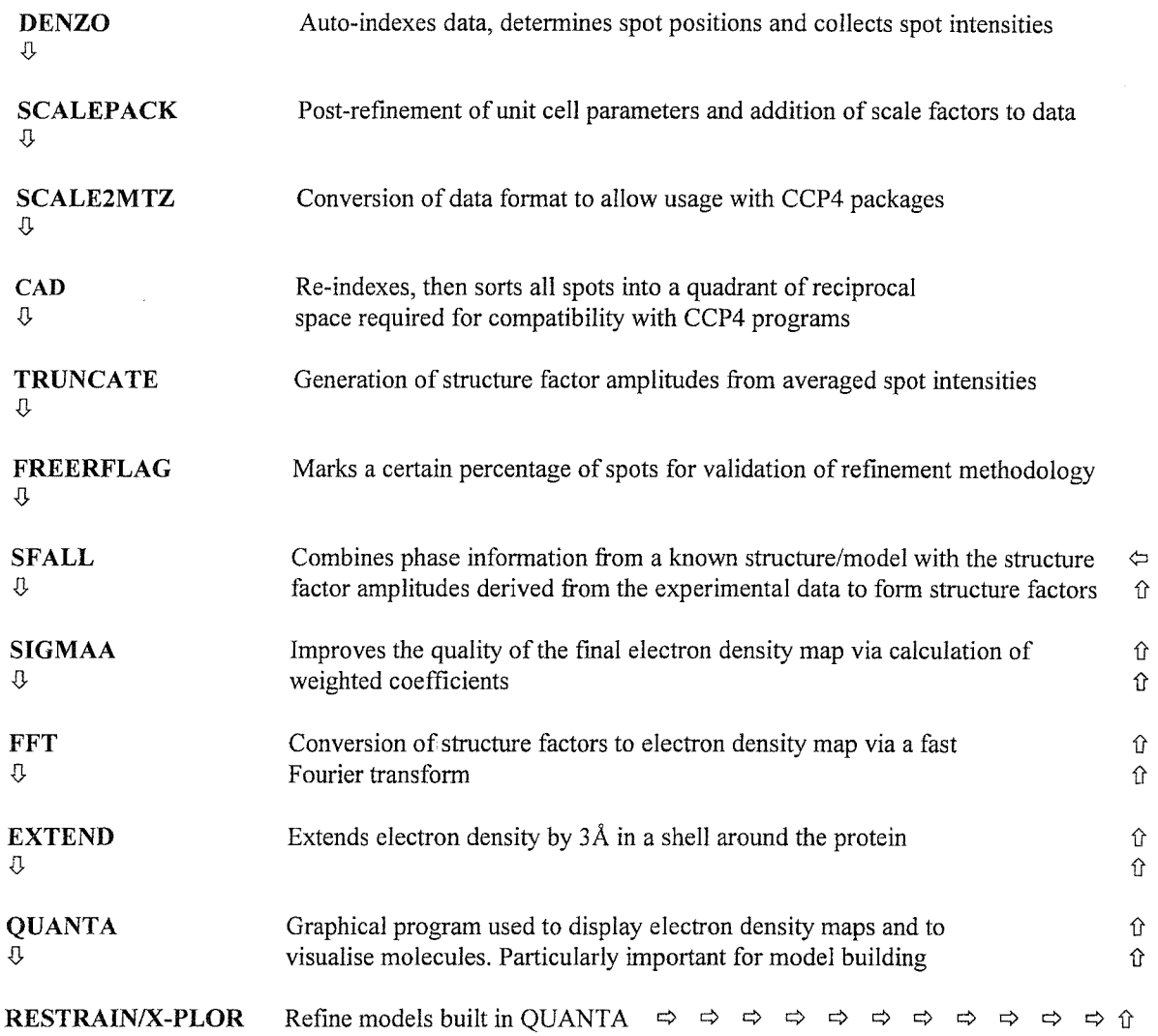


Figure 16(a). Summary diagram displaying how all of the crystallographic computer programs used for the data processing of momordin-βNADH data inter-relate.

The data processing with momordin is a useful platform from which several of the key programs available to the X-ray crystallographer can be described in some detail. The overall scheme of programs used for this project is shown in figure 16(a) (above).

The first stage in any processing run involves the indexing of the spots seen in the diffraction data. This involves assigning each of them a unique three-digit number comprising the integers h, k and l. These integers are known as Miller Indices and represent coordinates

describing the position of the diffraction spots in a three-dimensional reciprocal-space lattice containing all of the possible diffraction spots from the crystal. A given hkl value also defines a particular set of invisible diffraction planes within the crystal which give rise to a corresponding diffraction spot (i.e. with the same h, k, l value). The planes exist as sub-divisions of the unit cell and their spacing is expressed in whole fractions of the cell dimensions as determined by the Miller indices.

Historically, indexing of spots was carried out by first orienting the crystal manually, and then, via precession pictures, assigning h, k and l values depending on the spot positions relative to the origin of the reciprocal lattice and to each other. This method has now, however, been superseded by **auto**-indexing, a method employed by all current initial data processing packages which provides computational assignment of Miller indices to spots at any crystal orientation. As with all methods of indexing, auto-indexing requires a knowledge of the unit cell dimensions and space group (the latter describes the crystallographic symmetry between protein molecules present in the unit cell which arises from centring, rotation- and screw axes). From these we can obtain predicted positions for all of the possible spots in the reciprocal-space lattice for the crystal under investigation.

The momordin-βNADH data was auto-indexed using DENZO [165]. This software package can auto-index the diffraction data using a single diffraction pattern. The program performs a full real-space vector search in order to determine the best set of independent vectors which will allow indexing of all the strong spots picked on an image. From this, a unit cell of minimal volume called the reduced cell can be determined, belonging to the triclinic crystal system where symmetry elements are at their absolute minimum. The computer then distorts the reduced cell to fit each of the 14 Bravais lattices in several different ways and the predicted spots in each instance are compared with the pattern of strong spots in the diffraction pattern as before. A penalty is then applied to each of the unit cells found which

serves as a guide to how well each of the unit cells' predicted diffraction spots fit the observed pattern. The best cell to select is typically the Bravais lattice with the highest symmetry but with very low penalty, but it is sometimes best to resolve ambiguity by selecting the triclinic cell for an initial run of data processing to see at a later stage if additional symmetry elements are evident in the diffraction pattern, after which processing can then be re-run with the appropriate symmetry restrictions imposed.

Once a suitable Bravais lattice and cell dimensions have been chosen, additional symmetry elements still need to be considered in order to obtain the full picture describing the unit cell with all of its symmetry elements. These are rotation- and screw axes. Once again, with initial processing, the symmetry should be kept to a minimum since whether any is actually present is not known until after data processing when all of the spots have been collected and whose symmetry effects show up as mirror-images and systematic absences within the full data spread. Ultimately then it is not uncommon for multiple data processing runs to be carried out for protein crystals in order to improve confidence in the choice of space group.

After auto-indexing, the predicted spot positions need to be refined so that they accurately pinpoint all the spots in the reciprocal-space lattice and do so in such a way as to ensure all of the diffracted intensities are fully collected by the computer. The refinement involves the adjustment of a number of physical parameters relating to the crystal, beam and detector such as the beam centre, crystal orientation, unit cell lengths and angles, crystal-to-detector distance and beam divergence. If these parameters are correctly refined, then the spot predictions will match the spots more and more closely until the data is predicted optimally. DENZO deals with refinement in a user-controlled step-wise manner. The most important parameters, such as crystal orientation, tend to be refined first, and at low resolution, whilst most or all of the other relevant parameters are refined second and at high resolution. The parameter and resolution selections are under full user-control. To avoid problems associated

with highly correlated parameters, i.e. where increasing the errors in certain parameters can partly-compensate for the errors in others and in so-doing give the appearance of refinement, DENZO applies eigenvalue filtering, a method employed to remove the most correlated components from the refinement and make it numerically stable. DENZO also uses partials in the refinement to overcome problems associated with auto-indexing on a single diffraction pattern (i.e. one crystal rotation parameter is undefined, whilst the others are highly correlated and/or poorly defined), where the intensity of the partials is compared to the predicted partiality multiplied by the average intensity in the same resolution range. It should be briefly mentioned at this point that it is important during refinement to have a good estimate of the mosaic spread of the data and of the spot shape in three-dimensions. The mosaic spread is important for ensuring the optimum number of spots are identified and refined whilst the spot shape is critical for defining the boundaries of the spots, ensuring they are all treated as distinct points in the data during refinement. Optimal values for these two parameters in DENZO are chosen by the user based on output from the refinement log file.

The final parameter to optimise at this stage, when using DENZO, is the profile-fitting radius which defines an area around each spot, ideally containing between 10 to 50 additional strong spots. The spot profiles (i.e. three-dimensional shapes) within this radial distance from a given spot are averaged together and this is then used to define the specific profile of the spot at the origin of the radius which is to be used in intensity integration.

The end product of positional refinement on one diffraction pattern is the precise matching-up of predictions over spots throughout the image. In DENZO, the markers representing the degree of matching up are the χ^2 values, where a successful goodness-of-fit value is shown as a χ^2 value of 2.0 or less. Once this level of agreement between predictions and spots has been reached, several (typically 10-20) images are then examined to see if the refined parameters for one image may be generally applied to the whole of the collected data. With high

symmetry crystals (trigonal, tetragonal and cubic), the next ten successive images are normally chosen to test the positional refinement, whereas with low symmetry crystals (triclinic, monoclinic and orthorhombic), two wedges, each of ten images and 90° apart, are chosen to check the positional refinement.

Once the prediction-spot match is optimised for several images, integration (i.e. collection) of the spot intensity information from the full data set can then finally proceed, where intensities are first collected at all identified spot positions and then corrected via background intensity subtraction and application of Lorentz, polarisation and absorption corrections. With DENZO, the background value for a particular spot arises from an averaged intensity of the background pixels in the immediate vicinity of the spot.

Once a suitable number of spot intensities have been integrated, the next task is to perform post-refinement on the cell dimensions and crystal orientation, to scale all of the images together and to combine symmetry-related spots (otherwise known as multiplicity or redundancy in the data) and thence determine R_{merge} values from these. The R_{merge} values represent a measure of the accuracy of the integration and scaling steps. Post-refinement is required at this stage particularly to provide ideal unit cell parameters, since these will vary over the full range of images in a data set due to lack of clarity of certain cell dimensions in particular images.

Scaling is a crucial process where fluctuations in several parameters from one image to the next are all smoothened out, resulting in a series of diffraction patterns whose key fluctuating parameters are all made relative to one another, and thus, the collected intensities from one image to the next in the data, now form part of a fully-scaled data set, rather than being 'independent' from those on adjacent images.

The scaling applies a scale factor to each image to correct for fluctuations in parameters such as beam intensity (especially relevant when using a synchrotron source to collect data), radiation damage to the crystal, and absorption.

The way in which scaling is achieved is by matching up symmetry-related spots from different images which should have the same intensity. Any variation in intensity between symmetry-related spots can be taken into account via applying a scale factor to either of the intensities in a pair of symmetry-related spots, to make their values uniform, i.e. this provides information on how the strength of diffraction varies from one image to another. This principle is repeated for thousands of symmetry-related spots in the data which each contribute to the scale factors for the images which they lie on. The final result of all of this then is the aforementioned net scale factors, one of which is applied to each diffraction image. It should also be noted that a B-factor value for each of the images may also be estimated. The B-, or temperature-, factor, is a numerical value associated with the flexibility of atoms within a crystal. In scaling it is used largely to account for radiation damage in the crystal.

The final stage in scaling is the merging together of data and calculation of R_{merge} values. This involves determining average intensities for symmetry-related spots **after** scale factors have been applied to the data, and then retaining these merged intensities for future manipulations rather than keeping all of the independent intensities obtained for each of the symmetry-related spots. The R_{merge} value is a record of this intensity deviation between all of the symmetry-related spots, and is thus a guide to how well the symmetry-related spot intensities agree with one another. Consequently, the R_{merge} value is a useful measure of data quality.

The momordin data was scaled using the algorithm of Fox and Holmes [166] in the program SCALEPACK [165].

The bulk of the subsequent software used for this project came from the CCP4 suite of programs written for the field of X-ray crystallography [167]. Some of these programs will now be briefly discussed.

After scaling, the next (CCP4) program encountered is SCALE2MTZ which simply converts the scaled, merged data file into a format readable by the CCP4 programs.

CAD was the next program run. This re-indexes all of the collected spots and then places the merged spots in a particular region of reciprocal space which is necessary for input into other CCP4 programs.

Finally, TRUNCATE was run. This converted each spot intensity into an all-important structure factor amplitude.

Next up was the CCP4 program FREERFLAG which was used to set aside a random set of spots (5% of the data was selected in this case), which is kept apart from the rest of the data and not used directly in the later model refinement, but which is used separately at that stage to validate the refinement methodology (i.e. at a particular stage of refinement, even though the Rfactor may drop, if the Rfree is constant and/or increases in value, then the refinement should be stopped since it cannot proceed any further with the data available).

The next four CCP4 programs encountered with the momordin data are crucial to all X-ray structure determinations. First there is SFALL which, in this example, combines phase information from a known structure (in this case, native momordin from the Protein Data Bank [37]), with the structure factor amplitude information present in the experimental data (momordin with ligand present). The result of this is the generation of the all-important structure factors.

Next up is the program SIGMAA. This program calculates weighted coefficients which are then applied to the data, the result of which is an improvement in the quality of the final electron density map.

Next we come to FFT which, via a fast Fourier transform, converts the structure factors into an electron density map. The program EXTEND then, as the name might suggest, extends the electron density map around the protein molecule in three-dimensions, and thus the electron density map is ready for input into a graphical display package.

The graphics package used for the proteins studied throughout this work was QUANTA [168]. This allowed modelling of the βNADH-bound momordin into the electron density map generated above.

Several forms of electron density map may be fed into QUANTA to assist with modelling, e.g. the 2Fo-Fc map which provides a full electron density chain spanning the full length of the protein (where "Fo" represents information from the experimental data and "Fc" represents information from the model). Next, there is the positively-contoured Fo-Fc map which reveals regions of electron density present in the experimental data but which have no atoms modelled into them (e.g. where a side chain may have been re-oriented in the momordin-βNADH structure, a region of electron density in the shape of the side chain, and projecting out into space, would be present in this electron density map). Conversely, there is the negatively-contoured Fo-Fc map which shows regions of electron density in which there should be no atoms present in the structure (i.e. where there is no information from the experimental data). Using these three electron density maps then, the new structure can be modelled. There is one final process required however, before the momordin-βNADH structure can be solved. This is the refinement of the model to ensure all atomic separations, temperature factors, bond lengths and angles are optimised. With the momordin data, a combination of the CCP4 program RESTRAIN [169] and X-PLOR [170] were used for the refinement. RESTRAIN refines the model using the so-called least-squares method which has a limited radius of convergence. The program can be run with either constraints (as with rigid body refinement where large protein units, such as domains, can be refined), with restraints

(as with refinement of small regions of the structure, where stereochemical parameters are allowed to vary around a standard value, controlled by a weight), or with a combination of the two.

X-PLOR achieves refinement using so-called conjugate gradient refinement, which is again based on least squares methodology. This can provide quicker convergence than traditional least squares methods. Furthermore, rather than using geometric restraints in the refinement, X-PLOR uses an experimentally-derived potential energy function. X-PLOR can also use simulated annealing for the refinement. This is where the program simulates the super-heating of all of the protein atoms followed by a gradual cooling until all of the atoms are in their lowest energy state. This avoids the problem of the refinement becoming trapped in a local minimum by allowing uphill as well as downhill search directions.

Another key component of the refinement process is the Rfactor. This is a numerical representation of the agreement between the observed (experimental) and calculated (model) structure factor amplitudes and is defined as:

$$\text{Rfactor} = \frac{\sum_{hkl} \big| \, |Fobs| - k|Fcalc| \, \big|}{\sum_{hkl} |Fobs|} \times 100\% \qquad \text{Equation 2.}$$

where $|Fobs|$ = observed structure factor amplitudes, $|Fcalc|$ = calculated structure factor amplitudes and k = a scaling constant (puts $|Fobs|$ and $|Fcalc|$ on the same scale).

Starting models tend to have high Rfactors (e.g. 40-50%), but as refinement proceeds and the observed and calculated amplitudes come into closer agreement, the Rfactor generally drops to a value of around 20%.

A serious drawback with the Rfactor is that it can be inadvertently adjusted to too low a level. This is because the model can be over-fitted to the data, i.e. incorporation of excess water molecules, manipulation of temperature factors, use of restraints rather than constraints, etc., can lead to the generation of a model that has minor errors incorporated into it which go undetected.

As a result, the aforementioned free Rfactor (Rfree) was introduced for cross-validation of the refinement methodology [171]. Once again, this is used where a sub-set of randomly chosen spots are omitted from the refinement and tested to see how well the model predicts them. The Rfree is a numerical representation of this process. Since the spots involved in Rfree calculation have been left out of the refinement, they give an unbiased indication of the success of the refinement. Once refinement has been optimised, the test-set is then re-integrated into the last cycle of refinement.

As can be seen from figure 16(a) (above), modelling and refinement together represent a cyclic stage of structure determination. Thus, at this final step in the overall process, the electron density fits the experimental data more and more closely, until the point is ultimately reached where the match is at its best, at which point the structure has finally been solved.

2.3.3 Processing of selenomethionine MhpC multiwavelength anomalous diffraction data

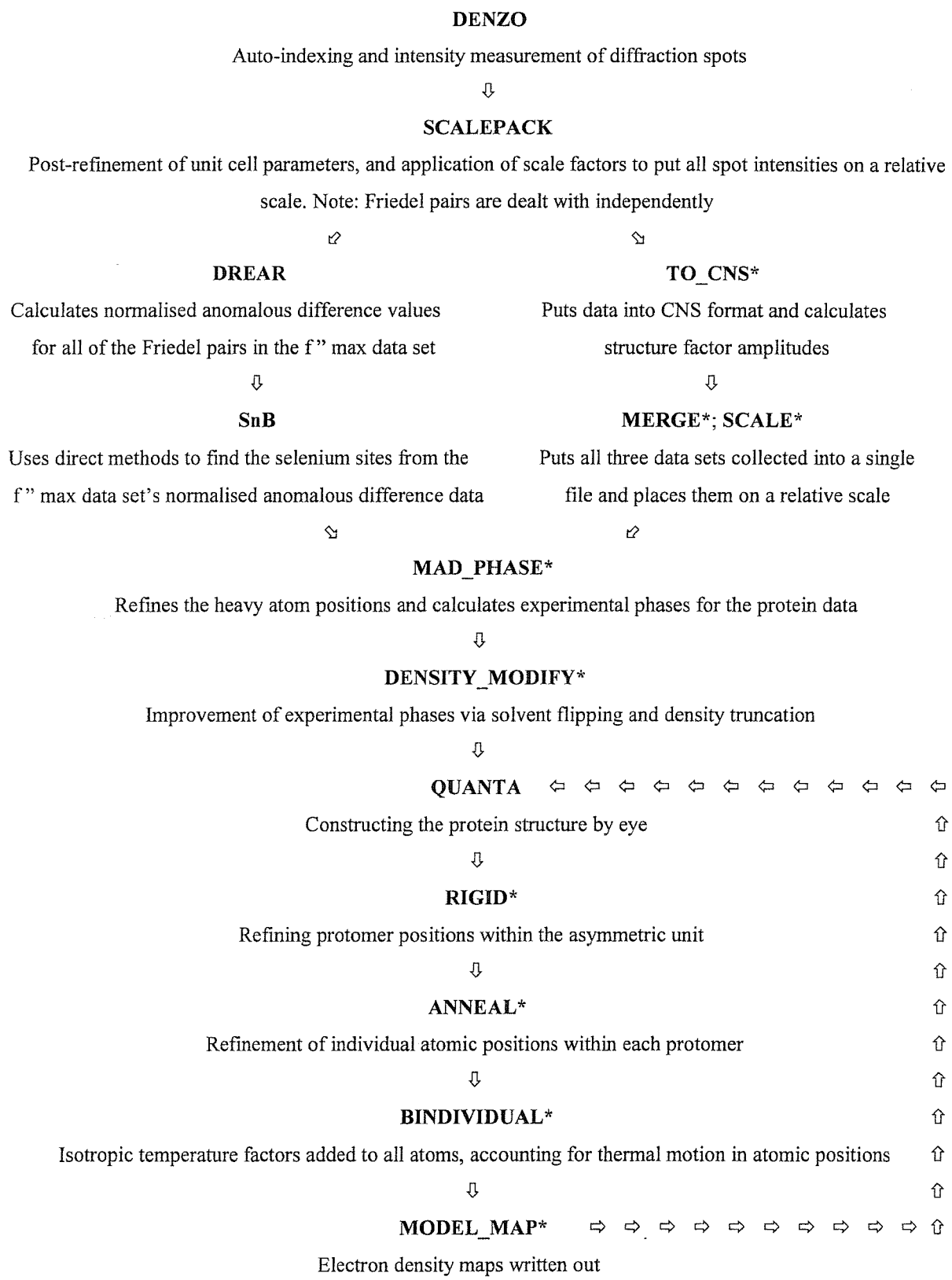


Figure 16(b). Summary diagram highlighting the principal computer programs used in generating the MhpC structure. Note: * indicates programs derived from the software package CNS [172].

2.3.3.1 The Theory of M.A.D. Phasing

M.A.D., or **M**ultiwavelength **A**nomalous **D**iffraction (or **D**ispersion) phasing is a special technique employed by the X-ray crystallographer to solve the infamous phase problem. The method has become increasingly popular within the X-ray crystallographic community over the past decade as more and more advances have arisen which favour its use. Principally, wavelength-tuneable synchrotron X-ray sources have emerged in greater abundance [173], which, as we shall see, represent a fundamental requirement of this technique, and in addition, developments in cryocrystallography [173], (CCD) detector technology [173] and in data collection strategy programs have allowed extremely rapid data collection of multiple data sets to proceed on a single crystal.

Application of the technique to a given problem requires that two criteria be met: firstly, an electron dense element, typically of atomic number 26 or greater [174], must be present within target protein crystals. This element may reside naturally within a protein, for example, calcium, zinc, iron or copper ions [175], or alternatively it may have to be introduced into a protein by an experimenter, for example, selenium which may be incorporated into amino acids [175], xenon, which may bind to proteins under high pressure [173], or mercury, which can be soaked into crystals [175]. The second criterion necessary for a M.A.D. phasing experiment is, as mentioned above, having access to a wavelength-tuneable X-ray source. The importance of these two factors should become clear from the text that follows.

The theory behind the M.A.D. phasing technique will now be presented. Normally, in an X-ray crystallographic experiment, pairs of diffraction spots, termed Friedel mates, arising from the same set of crystallographic planes, have the same intensity. In a M.A.D. phasing experiment however, these pairs of Friedel mates have **different** intensities to one another. This unusual observation is due to a phenomenon called **anomalous scattering** [175].

At the atomic level, anomalous scattering occurs when X-rays bombarding a target crystal are at a specific wavelength, and corresponding energy, which is equivalent to that energy required to boost an electron out of an atomic orbital [175]. When this situation arises, resonance is set up in the system, whereby excited electrons gain vibrational energy from the incident X-rays. As a consequence of this energy transfer, the X-rays themselves are effectively absorbed. After a very short time, the excited, vibrating electrons return to lower energy shells within their atoms, and in the process, re-emit the X-rays, but with altered phase [156]. It is this phase variation between incident and re-emitted X-rays which accounts for the anomalous scattering mentioned above.

The precise energy of the X-rays at which anomalous scattering occurs is element-dependent and is termed the **absorption edge** [175]. Absorption edges for light atoms such as carbon, nitrogen and oxygen in proteins are not near the X-ray energies used in X-ray diffraction studies, and as such, these atoms do not give rise to a significant anomalous signal in crystallographic work [156]. However, heavier atoms have absorption edges which correspond to selectable X-ray energies (available at synchrotron sources), and as such, a given 'heavy' element can give rise to an anomalous signal when an X-ray energy matching up with the 'heavy' element's absorption edge is used [156]. Such an anomalous signal then is effectively solely due to the specific 'heavy' atoms present inside a target molecule. Therefore, the anomalous diffraction data in a M.A.D. phasing experiment, which, once again, takes the form of the intensity differences between Friedel mates, can provide information about the heavy atoms **alone**. In fact from measuring such intensity differences, statistical relationships among them allow the phase problem for the just the heavy atoms to be directly solved, i.e. the heavy atom positions can be found. Once the wave phase and amplitude contribution to each diffraction spot due to the heavy atoms is known, then the phase information for the protein alone can finally be obtained [156].

The M.A.D. phasing technique has certain benefits over other phasing methods. Firstly, since only one crystal is normally used for complete data collection, then all M.A.D. data sets collected are typically free from anisomorphism, i.e. they are devoid of distortions between the same spots seen in different data sets collected at different wavelengths which arise as a consequence of physical differences between different crystals. Anisomorphism leads to a reduction in overall data quality, introducing, for example, errors associated with scaling and merging of data sets, which are potentially damaging in M.A.D. phasing experiments since the anomalous data which provides the phase information, represents only a small fraction typically 3-8% of the overall diffraction, and as such, intensity measurements need to be highly accurate [173]. Thus, the benefits of using a single crystal for full data collection are quite rewarding in terms of final overall quality of the resultant electron density maps.

A second benefit normally associated with the M.A.D. phasing technique is absolute knowledge of which amino acids in a target protein contain the anomalous scatterers. However, this is only true for certain anomalous scatterers which can be physically incorporated into a specific amino acid within a protein, such as selenium, which can be specifically introduced in the form of selenomethionine into proteins in place of native methionine residues. This is useful when initially assigning amino acids to suitable regions of electron density, where the primary sequence of the protein is known. This is because knowledge of where the anomalous scatterers are, pinpoints precisely where the amino acids housing them are in the protein structure, and thus, building up the rest of the protein around these positions is relatively straight forward, using the amino acid sequence of the protein as a guide. Partly because of this advantage, selenium in the form of selenomethionine is most often the choice of anomalous scatterer used in M.A.D. phasing work. It also does not disrupt the native structure of proteins it is incorporated into [175], and in the form of selenomethionine, can normally become incorporated into proteins in low quantity since

methionine's natural abundance in proteins is around 2% [175] (compared with the expected average of 5%), which makes solving of phases far simpler and therefore more likely to succeed.

The M.A.D. phasing technique is not without its limitations. For example, data will be uninterpretable with too many anomalous scatterers inside a protein. This generally means that large protein structures will not be solvable using this technique due to the occurrence of excessive numbers of anomalous scatterer sites inside such proteins. Secondly, where selenium is desired as the anomalous scatterer in the form of selenomethionine, proteins totally deficient of methionine residues cannot be used for the M.A.D. phasing technique. However, this problem can be overcome via mutagenesis of typically leucine, isoleucine or valine residues to methionine. These residues are the most frequent replacements for methionines in natural proteins [173].

2.3.3.2 Data Processing – Selenomethionine MhpC

Multiwavelength anomalous diffraction data on selenomethionine MhpC crystals were collected at three wavelengths at the European Synchrotron Radiation Facility, Grenoble, France. The details of the processing of this data are described below. The overall scheme detailing the principal programs used is shown in figure 16(b) (above).

Initial data processing of the three collected data sets employed the CCP4 programs: MOSFLM, SORTMTZ, SCALA, TRUNCATE and HKLVIEW [167]. Descriptions of these programs follow.

MOSFLM is an equivalent program to DENZO in that it allows auto-indexing of diffraction data and performs subsequent positional refinement and intensity integration of the spots, all using an interactive image-displaying window. A new fast Fourier transform-based auto-indexing algorithm [176] has recently been introduced into MOSFLM and was used for the

processing of the three data sets. The program finds the spot positions from the data and then uses them to compute the corresponding reciprocal-space lattice vectors. The vectors are then projected onto a chosen direction and are subjected to a one-dimensional Fourier analysis. As a result, a large Fourier term is found which has a periodicity corresponding to the interplanar distance, if the chosen direction is perpendicular to a set of widely separated reciprocal-lattice planes. Identification of such directions over a hemisphere establishes the best potential basis vectors of the real cell [176].

During auto-indexing, a distortion index table is presented, as with DENZO, and once again, the best choice of Bravais lattice from this table is most likely to be that with the highest symmetry, but with low penalty (typically below 10 here). Once the Bravais lattice has been selected, the preliminary space group for processing is next chosen from the limited list possible for a given Bravais lattice type. The space group defines the crystallographic symmetry from centring, rotation- and screw axes, which determines where protein molecules are found within the unit cell. With initial processing, the optimal choice of space group is that with the lowest symmetry, since this will allow clear identification of the most probable space group at a later stage during processing (i.e. at the HKLVIEW stage. See below).

After auto-indexing, the next steps are the refinement of the detector, beam and crystal parameters which allow optimisation of the fit between the collected diffraction spots and their corresponding predicted positions. Whereas with DENZO, where the parameters to be refined and their order of refinement are under user control, with MOSFLM this is a more automated process. Detector parameters are refined using least squares methodology where reflections are initially picked from the centre of an image and then from over the entire image during the refinement process. The default is to just use fully recorded spots for the refinement, but partially recorded spots can also be used where there is insufficient fully recorded data.

Whereas in DENZO the markers for a successful positional refinement are the χ^2 values, with MOSFLM, the equivalent values are the positional residuals. These are measured in millimetres, where the maximum allowable discrepancy in distance between a given spot and its predicted position is determined by the diffraction pattern pixel size.

After initial refinement of the positional parameters has been achieved, the size of the spot measurement boxes is next optimised, and then post-refinement of the crystal and beam parameters is performed.

The overall refinement process uses several images from the diffraction data. Once again, low symmetry lattice types (orthorhombic, monoclinic and triclinic) typically use two segments of data in refinement, which are widely separated from one another, and each of which commonly comprises 5-10 degrees of data in phi space for a successful refinement. Two wedges of data are needed due to the lack of symmetry within the full data set, which means that a lot of diffraction images must be processed, and thus parameters must be refined so that predictions overlay spots throughout the data. In contrast, high symmetry lattices (trigonal, tetragonal and cubic) may only need a pair of images from one segment of the overall data for refinement. This is because the symmetry in such data is high and so only a relatively small fraction needs to be processed, and thus the predictions refined using a pair of images should reliably overlay the spot positions in all images when processing so few diffraction patterns.

Once the spot positions and general sizes have been determined, the intensities of the individual diffraction spots must next be gathered. This involves the generation of standard profiles for each spot for profile fitting. Each diffraction pattern is divided up into a grid of rectangles of equal area (nine rectangles for low resolution data and twenty-one for high resolution data), and fully-recorded spots (in general) above a certain threshold value in each rectangle are then used to calculate a standard spot profile for that particular area of the detector. Profile-fitted intensities and standard deviations are then obtained for each spot

using weighted profile-fitting methods: For each spot, a new profile is determined by linear interpolation of the standard profiles of the areas surrounding that spot. The interpolation is based on the distance of the spot from each standard profile, where the boundary of the spot profile being determined is found from the intensity-weighted mean of all the spots contributing to that profile. This method represents an accurate way of modelling how the spot profiles vary across the detector face. The summation integration intensities and associated standard deviations are also determined for all of the diffraction spots. These are alternative measurements of the spot intensities and standard deviations, for comparison [177]. Finally, a matrix is calculated which defines the unit cell lengths and angles along with the crystal orientation information. (It should be noted that the matrix found for one of the three data sets collected in this work was used as the initial matrix for each of the other data sets. However, with the other two data sets, the matrix was slightly refined to fit each of these precisely before processing was performed).

The data processing thus far described is all preliminary work based on a few of the collected images, which is necessary to define reasonably accurate physical parameters defining the nature of, and interaction between, the crystal, X-ray beam and detector. Once these parameters have been defined, the full data set may then be processed, whereby all of the diffraction spot positions are catalogued and their intensities measured.

With all of the spot intensities collected, these are then partly adjusted using the Lorentz, polarisation and absorption corrections, and the list of spots is then placed in numerical order according to the values of the Miller indices. This is achieved by the CCP4 program SORTMTZ. Next, the program SCALA applies scale factors to the data using a similar algorithm to that employed by SCALEPACK [FH], and it then merges the data, outputting the all-important R_{merge} statistics. It should be noted that in order to retain the anomalous differences between Friedel pairs in the data, the Friedel pairs are not merged after scaling.

Finally, the scaled and merged data output from SCALA is fed into TRUNCATE where the number of molecules per unit cell is estimated, and structure factor amplitudes are calculated. The output from TRUNCATE is then used by the CCP4 program HKLVIEW to confirm the choice of space group, via viewing pseudo-precession pictures of all of the collected spots in a particular data set, and by looking for symmetry elements and systematic absences in the data. Once this stage is reached, if the space group initially chosen during auto-indexing is found to be incorrect, then processing must be fully repeated using the new space group. With this particular project, the program path just described was used to identify the space group, but it was then found that the output from TRUNCATE was incompatible with the direct methods search program SnB [178,179,180] which was next to be used to locate the anomalous scatterers. Thus, the space group, and other information found from this program path, was thence used to reprocess the diffraction data using the SnB-compatible programs DENZO and SCALEPACK [165].

For each of the three data sets collected, DENZO [165] was used to re-perform auto-indexing, to help verify the choice of space group, to once again refine the diffraction spot positions and to calculate the spot intensities. The program was used in the same way as with the momordin-βNADH data, and as such, for a description of how DENZO functions, please turn to section 2.3.2.

Once all of the diffraction spot positions and intensities were determined, each of the three data sets was then independently scaled using the Fox and Holmes algorithm [FH] once again in SCALEPACK [165], and the unit cell parameters were also refined. The details of how SCALEPACK works are described in section 2.3.2. However, scaling is performed slightly differently where anomalous data is present since the Friedel pairs do not share the same intensity values in this instance, and so cannot be used to scale one image to another.

However, scaling can be achieved using other groups of symmetry-related spots which are still expected to have identical intensities. These are termed Bijvoet pairs.

The presence of Bijvoet pairs is dependent on symmetry between protein molecules within the crystal unit cell, where higher symmetry gives rise to more Bijvoet pairs.

It is important to note at this stage that during scaling of anomalous diffraction data, since the intensity difference between pairs of Friedel mates needs to be measured in order to solve the phase problem, then the Friedel pair intensities must be kept as independent values, i.e. they must not be averaged and merged together, although their intensities may clearly be adjusted slightly due to the scaling process. In SCALEPACK, the command line "scale anomalous" is used in the input file to specify that the Friedel mates keep their individual intensities.

After scaling, the so-called f" max data set, which was collected at the peak of the MhpC selenium's absorption edge, where the anomalous signal was maximal, was fed into a program called SnB [178,179,180]. SnB was employed to find the position of all the anomalous scatterer sites within a structure. This program was chosen since it deals effectively with a large number of anomalous scatterers, which was beneficial to this particular problem since there were 32 selenium sites predicted to be in the asymmetric unit and which thus had to be found. This large number of anomalous scatterers represents a very challenging problem, hence the need for an automated search program.

The SnB software package includes an additional program called DREAR [181], which is run as a necessary precursor step before SnB can be used. The DREAR program measures all of the anomalous intensity differences between Friedel mates in the f" max data set, and it then simply calculates normalised anomalous difference values for each of these measured differences.

The anomalous data is now in a format suitable for input into SnB. SnB uses direct (statistical) methods to find the anomalous scatterer (i.e. selenium) positions. The program

selects random phases that aim to represent those of the net diffracted X-rays emerging from the crystal, which are due exclusively to the presence of the heavy atoms. These phases are then combined with the normalised anomalous difference intensity information derived solely from the anomalous scatterers present in the crystal. The experimentally-derived amplitudes are used to refine the phases and these can then be used to generate an electron density map upon applying a fast Fourier transform to them. The resultant electron density map may then be modified by the SnB program to see if the random phases it has used can give rise to realistic selenium sites or not. This procedure is repeated using, typically, several hundred different sets of random phases, until a set is found that tallies well with the normalised anomalous experimental data, and which thus gives rise to well resolved anomalous scatterer sites.

Many potential solutions for the positions of the anomalous scatterer sites are output by SnB, and these need to be tested for their validity. The way in which the solutions were checked with the selenomethionine MhpC data was using a CCP4 program called FINDNCS. This takes a set of likely anomalous scatterer positions output from SnB, and then tries to rotate groups of these sites around various positions in the unit cell to see whether they could be made to accurately overlap the positions of some of the other groups of sites found in the same set, i.e. the presence of high symmetry between groups of sites is sought after. Where this occurred with the selenomethionine MhpC data, it was a very strong indicator that the clusters of sites related by symmetry to one another were real anomalous scatterer positions within the unit cell. This is because each of these clusters of sites is arranged in three-dimensional space in an identical pattern to each of the others, which is very strong evidence suggesting that they are adopting non-random conformations.

Once the selenium sites had been found, the next task was to obtain estimates of the experimental phases for the protein data. For this stage of the data processing, a database of

crystallographic (and nuclear magnetic resonance spectroscopy) programs called CNS [172] was used. Firstly, the CNS program TO_CNS was used to put the scaled data collected at all three wavelengths output by SCALEPACK into CNS format, and to also calculate structure factor amplitudes corresponding to each of the scaled diffraction spots. Next, CNS programs MERGE and SCALE placed all three data sets into a single file, and then put them all on a relative scale.

The data from all three wavelengths was now in a state suitable for estimation of the experimental protein phases. The CNS program used for this procedure was MAD_PHASE. This software takes the selenium site positions output by SnB and refines them using both the anomalous differences between Friedel pairs in the f " max data set, and also using the dispersive (intensity) differences between identical spots found in the other two data sets, termed the "f ' min" and "remote", each of which was collected at a unique X-ray wavelength that gives rise to a different degree of anomalous scattering in the data. From the refined selenium positions, more accurate phase information can be obtained, which represents their contribution to the overall scattering collected in the diffraction patterns. MAD_PHASE then takes this information, and also the structure factor amplitude information for each spot in each of the three data sets, and from all of this data, the program is able to calculate estimates for the experimental protein phases.

Once MAD_PHASE has been run, an additional problem must next be addressed before moving on. This problem arises due to the unknown handedness of the anomalous scatterer sites, i.e. two enantiomeric arrangements of the anomalous scatterer sites are possible from the single solution output by SnB, and as such, only the phases output from the correct enantiomeric arrangement will give rise to the correct protein structure. Thus, a CNS program called "MAD_PHASE_FLIP" must be run using the second enantiomeric arrangement of the anomalous scatterer sites to generate estimated protein phases using the same methods as

employed by MAD_PHASE. Once both sets of experimental phases are available from both arrangements of the anomalous scatterer sites, then electron density maps may be calculated using both of the sets of experimental phases. The correct solution should produce a map in which density is non-uniform, and where distinct boundaries can be seen between protein electron density and background solvent.

Once the hand ambiguity is resolved, the CNS program MATTHEWS_COEF is next used. This provides an estimate of the solvent content of the crystal. This is achieved by first predicting how many copies of each protomer will be present in the crystal unit cell, then by estimating their collective volume, and finally by subtracting this from the total unit cell volume. The estimated solvent content is used in solvent flipping/flattening and in density truncation which are techniques employed to provide better definition of both the protein and solvent in the next CNS program run, DENSITY_MODIFY, which thus helps to improve the electron density map at this stage.

DENSITY_MODIFY takes the initial estimates of the protein phase information output from the MAD_PHASE program, and then modifies these. It also estimates structure factor amplitudes and phases for any diffraction spots missing in the collected data, and in so doing helps to provide an improved electron density model of the protein. The map output from DENSITY_MODIFY is thus used to generate a preliminary model of the protein structure. This electron density map was output in "NEWEZD" format in the work presented. This needed to be converted to X-PLOR format in order to be fed into the graphical display package QUANTA. The program used to convert the map format was the CCP4 suite's mapman.

Once read by QUANTA, the electron density map could next be used to construct a primary model for the MhpC structure. The first stage in this process was to generate bones. Bones are simply an alternative way of representing tubular electron density as averaged strings of

density which thread precisely through the middle of the electron clouds viewed on the graphical display. The bones simply act as a separate visual aid to assist with correct protein chain assembly. They are particularly useful at clearly distinguishing between side chain and main chain density.

The next task is to assign backbone α-carbons to positions at regularly spaced intervals along the electron density chain. The selenium sites pinpoint the methionine side chains within the structure very clearly, and as such can be useful starting points for finding primary Cα positions to build away from. Knowledge of the primary structure of the protein also helps to decide where each of the side chains branch off away from the main chain, especially where gaps in the electron density are present. A secondary structure prediction is also useful to have, since this may help you build in the Cα positions correctly in sequence by looking for secondary structural features in the electron density which you are building into.

Once all of the Cα positions have been modelled into the electron density, these are next labelled according to the specific amino acid types which they represent. QUANTA has an option to incorporate an entire amino acid one letter code sequence for a protein under study into its program. Once this has been achieved, a few residues are manually linked to positions within the protein sequence, and after a sufficient number have been located in this way, the program then takes over and automatically assigns amino acid types to all Cα atoms built into the electron density based on the primary sequence input. The program can then automatically introduce all of the main and side chain atoms into the model. These atoms are all built by the program into the strongest density available around their particular Cα atoms. As such, the side chains tend to be found lying in main chain density, and so the final task at this stage of the modelling is to re-orient all of these side chains out into their most viable regions of

electron density. Once completed, the primary model is finished. The final task is then to refine the initial model towards the completed protein structure.

For convenience, only one protomer was modelled, and hence output, using QUANTA at this stage, since the structures of all other protein molecules in the unit cell, and hence crystal could be modelled from this and then optimised during refinement. The CCP4 program PDBSET can generate models for all other molecules in the unit cell from the modelled protomer structure output from QUANTA, and also from the non-crystallographic symmetry operators (or rotation axes) output from the aforementioned FINDNCS, which define how the different anomalous scatterer sites, and hence protein units, are symmetrically arranged with respect to each other.

Once a model for all protein units within the unit cell was obtained, this was then fed back into the CNS package of crystallographic programs, for refinement. The initial programs run were used to generate the necessary basic files required for refinement to proceed. GENERATE_EASY puts the coordinate file for the protein output by pdbset into the CNS format, and also writes out a molecular topology file detailing structural information tailor-made for the protein under study such as descriptions of atoms, assignments of covalent bonds (connectivity) and bond angles [172]. Another important CNS program at this stage is MAKE_CV which sets aside a random proportion of the spot intensity data for cross-validation during refinement, i.e. this singled out data is necessary to test the refinement procedure and check that it is running successfully. The file used in conjunction with this program is the merged and scaled structure factor amplitude data file output from CNS containing structure factor amplitude information for all three wavelengths collected. The percentage of data chosen for inclusion in the test set was 5%.

The initial refinement process began with rigid body refinement, using the CNS program RIGID. This refines the protomer positions within the unit cell as separate rigid bodies, by

trying to optimise the fit of each of them to their local electron density. The atoms of a given protomer retain their precise orientation relative to one another during this refinement stage, and the procedure simply involves taking the whole of each protomer and rotating and/or translating it around as a rigid structure to best fit the density. Input required for this program is: the protein coordinate file in CNS format, from which the phase information is provided for refinement, the molecular topology file, the protein parameter file (which specifies various constants for conformational and non-bonded energy terms in proteins [172]), and finally, the structure factor amplitudes file (specifically the "remote" wavelength amplitudes) which also includes the data flagged for cross-validation (which was derived from the merged, scaled X-ray data file output by CNS). A refined coordinate file is output by RIGID. The next stage in refinement is performed using the CNS program ANNEAL. This performs simulated annealing, which is a simulation where all the atoms of the protein are superheated to a temperature of typically 5000°K and they are then gradually cooled in, for example, 50°K steps until the final thermodynamically stable structure is found at around room temperature. The annealing process is assisted by atomic restraints which dictate where certain atoms should roughly be with respect to their neighbours in the final structure. This program searches for the most energetically stable protein structure, and as such, it prevents the annealing structure getting stuck in false minima. In contrast to rigid body refinement, individual atoms are partially free to move around relative to all of the other atoms in the structure during simulated annealing. The input files required with this program are the same as those required for running RIGID, although an additional file is also needed here called NCS.DEF which tells the program which regions of the numerous protomers in the unit cell are symmetry related, and which also specifies the degree of flexibility that each atom to be refined has during refinement. A refined coordinate file for the protein is once again output. The third refinement stage is performed using the CNS program BINDIVIDUAL. This

applies isotropic temperature factors to all atoms in the structure, which account for the thermal motion seen in the atomic positions, i.e. application of temperature factors helps to rationalise regions of empty electron density in the structure by introducing information on atomic dynamics into the model. All of the input files required by this program are the same as those used for ANNEAL, including the NCS.DEF file. The first round of refinement is thus complete, and the resultant coordinate file is output. This is then split up into individual coordinate files, each providing coordinate information for a particular protomer.

Next, a CNS program called GET_NCS_MATRICES is used to determine the symmetry operators which describe the symmetry relating one protomer to each of the others in the unit cell in the refined structure. These operators are output to a new NCS.DEF file. Next, the CNS program MODEL_MAP is run, which calculates the 2Fo-Fc, positively-contoured Fo-Fc and negatively-contoured Fo-Fc electron density maps for the current model. The input for this program is the same as for RIGID, except that the newly generated NCS.DEF file is used to provide an electron density map for each protomer which is derived from averaged data from all of the molecules in the unit cell.

Once the electron density maps have been written out, a second round of model building in QUANTA can next proceed, using one protomer. After this, the model can then be applied to each of the other protomer positions in the unit cell using the CCP4 program LSQKAB. This takes the coordinate file of the newly modelled protomer and then rotates and/or translates this structure over the other protomer positions output and separated after BINDIVIDUAL. Finally, PDBSET is used to reassign chain identifiers (such as A, B, C and D) to all of the protomers present in the unit cell, and at the same time, the program is used to incorporate all of the protomers back into one single file again. After this, the resultant coordinate file is fed back into the CNS program GENERATE_EASY, which writes out the new coordinate file in CNS format once again, and which writes out a new molecular topology file, based on the

new model, and then refinement/modelling can proceed in cycles as described until the final structure is obtained.

Chapter 3

Materials and Methods

3. Materials and Methods:

All general reagents and chemicals were supplied by a reputable manufacturer. Where uncommon materials were used, a supplier has been mentioned.

3.1 *Der p*1 and asthma

3.1.1 *ProDer p*1 Fusion Protein Expression

*ProDer p*1 was engineered as a C-terminal fusion protein to *E. coli* His-patch-thioredoxin using the His-patch Thiofusion™ expression system (Invitrogen®)[182]. The *proDer p*1 gene was inserted into the pThioHisB plasmid, transformed into *E. coli* (strain: GI724) and cloned by our colleagues Professor Noor Kalsheker, *et al*, at the University Hospital, Queen's Medical Centre, in Nottingham. The cloned sequence of *proDer p*1 is shown in figure 17. A control plasmid, pThioHis B without insert, was also used as a positive control for expression. Cells were grown on LB agar plates and in LB medium cultures at both 30°C and 37°C. Induction of fusion- and control protein expression proceeded once the optical density of cells at 550nm reached 0.5-0.7. Isopropyl-β-D-thiogalactopyranoside (IPTG) was used as the inducer at a final concentration of 1mM. Cells were then harvested after 4-6 hours and spun down at 3000xg for 10 minutes at 4°C. Cells were resuspended in 20mM sodium phosphate with 0.5M sodium chloride, pH 7.8 and lysis occurred either via the French Press or Sonicator (used in conjunction with ethanol/dry ice snap freeze-thawing in-between cycles). Centrifugation then occurred at 3000xg for 15 minutes at 4°C, and both pellet and supernatant fractions were kept for determination of protein content. Protein expression was

monitored by Sodium Dodecyl Sulphate-Polyacrylamide Gel Electrophoresis (SDS-PAGE) using Pharmacia Biotech Phastgel® 8-25% gradient gels.

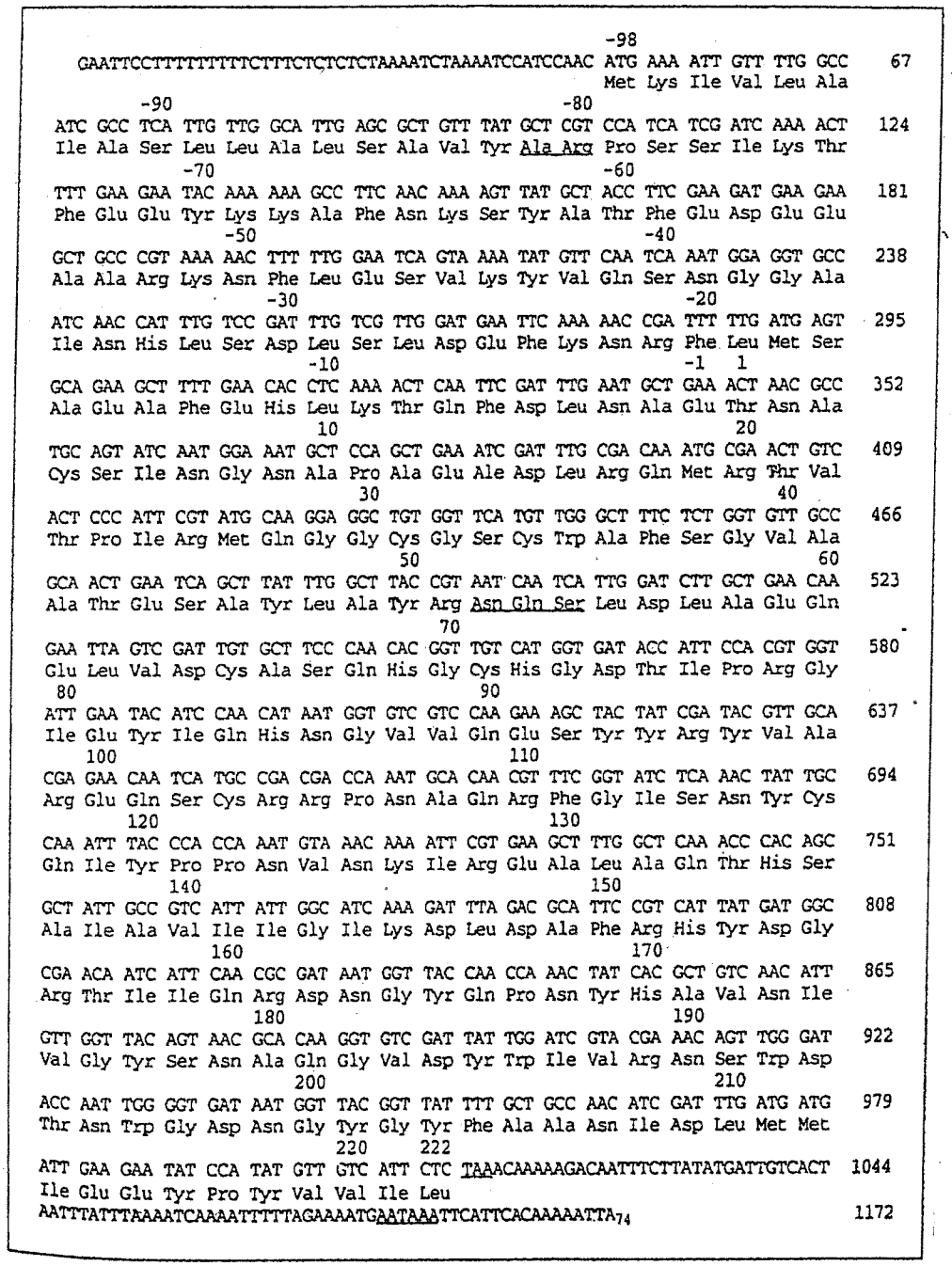


Figure 17. The cloned sequence of *proDer p*1 which agrees with the corrected sequence supplied by Chua, *et al* [19]. The precise region cloned spans from the codon encoding the upstream amino acid -80(Arg) to the codon encoding the downstream residue 222(Leu). Two extra cytosines were added to the 5' end of the insert to maintain the gene's open reading frame in the vector. The restriction enzymes *KpnI* and *XbaI* were used to insert the *proDer p*1 gene into the pThioHis B vector.

3.1.2 *ProDer p*1 Fusion Protein Washing-Refolding

Lysed-centrifuged cell pellets were washed to remove lipids in ice-cold RIPA buffer containing 0.1% SDS, 1% Triton X-100, 1% sodium deoxycholate, 0.15M NaCl, 1mM phenylmethylsulphonylfluoride (PMSF) and 0.025M Tris-HCl, pH7.5. The material was then vortexed and spun down at 5000xg for 5 minutes. The supernatant layer containing solubilised lipids was then removed and the wash repeated with the fresh addition of ice-cold RIPA buffer to the pellet. The inclusion body material was now ready for refolding.

The refolding protocol for propapain described by Taylor, *et al* [183], was followed for *proDer p*1. Solubilisation of protein pellets took place at 37°C over an hour in 0.1M Tris-acetate buffer, pH8.6, containing 6M guanidine-HCl (Gu-HCl), 10mM dithiothreitol (DTT) and 1mM ethylenediaminetetraacetic acid (EDTA). Solubilised material was thence diluted 10-fold in 0.5M Tris-acetate, pH8.6, containing 6MGu-HCl, 1mM EDTA and 0.1M oxidised glutathione and incubated overnight at 4°C. This mixture allows the formation of a protein-glutathione mixed-disulphide. The protein was thence refolded via 50-fold dilution in de-aerated 0.1M Tris-acetate, pH8.6, containing 0.4M L-Arginine and 3mM L-cysteine, which was also saturated with nitrogen gas, and subsequent incubation for 2 hours at 20°C. The sample was then concentrated using a YM-10 membrane and excess guanidine removed by dialysis in 50mM Tris-acetate, pH 9.0. SDS-PAGE was thence used on the resultant centrifuged-supernatant material in order to visualise soluble, refolded protein.

3.1.3 Autocatalysis of *proDer p*1

The pH of the soluble *proDer p*1 fraction was adjusted to pH4.0 with acetic acid, and cysteine was then added to the sample to a final concentration of 20mM. Aliquots were incubated at

either 40°C or 60°C for between 5-15 minutes and the occurrence of autocatalysis was checked using SDS-PAGE gels [183].

3.2 Momordin

3.2.1 Protein extraction and partial purification from seeds

Extraction and purification of momordin was carried out as per the method described by Husain, *et al* [127].

50g of seed material derived from the bitter gourd, *Momordica charantia*, was ground up in 300ml phosphate-buffered saline (0.1M sodium phosphate, 140mM sodium chloride, pH7.4). Lipids were then removed from the seed material by vacuum filtration over a Whatman 54 membrane using "Hyflo Supercel filter aid" (BDH). Further purification of momordin was then achieved by adding 60% ammonium sulphate to the filtered material, followed by centrifugation at 48,400xg for 30 minutes at 25°C. The resultant pellet was resuspended in, and then dialysed against, 50mM sodium phosphate buffer, pH6.3.

3.2.2 Gel filtration of momordin

The dialysed protein material was loaded onto a Sephadex G-50 gel filtration column and run with an optimal flow rate of 1.25 millilitres/minute. Determination of fractions containing momordin was by SDS-PAGE. Important fractions were pooled from several independent gel filtration runs and the material concentrated and gel filtered once more to improve the purity.

3.2.3 Cation exchange chromatography

Gel filtered fractions were pooled, buffer-exchanged into 20mM sodium acetate, pH4.5, and concentrated. The material was then loaded onto a sulphopropyl (SP)-sepharose-HP column and fractions collected using a 0-0.5M sodium chloride gradient. Momordin was eluted at 290mM sodium chloride concentration and fractions were pooled, buffer-exchanged into 20mM sodium phosphate, pH7.4 and concentrated to 15milligrams/millilitre(mg/ml).

3.2.4 Crystallisation of momordin

Momordin crystals grew up over two weeks at 17°C, using the hanging drop-vapour diffusion method (see below). Protein samples for crystallisation were at a concentration of 5-15mg/ml, and the mother liquor used for crystal growth consisted of 20mM sodium phosphate, pH7.4 along with 1-8% polyethylene glycol (PEG)-4000 or 2-10% PEG-8000. These conditions were adapted from those originally identified by Husain, *et al* [127].

The hanging drop-vapour diffusion method for protein crystallisation

Figure 18 shows a diagram of the basic hanging drop experimental arrangement.

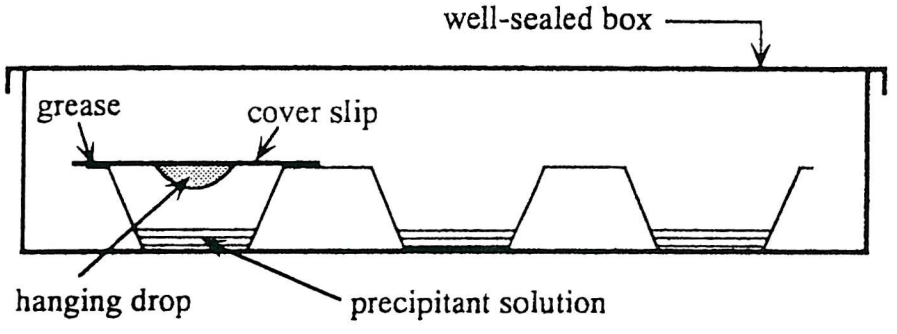


Figure 18. Diagram showing the basic experimental arrangement for the hanging drop-vapour diffusion method of protein crystallisation. A protein:mother liquor mixture hangs from a siliconised cover slip over a well of mother liquor. Crystals form in the drop as water vapour diffuses towards equilibrium between drop and well solution. Picture taken from "Principles of protein X-ray crystallography", Jan Drenth, 1994 [158].

This technique involves mixing, usually 2-4μl of concentrated protein sample (typically at a concentration greater than 1mg/ml) with an equal volume of a crystallisation mixture containing a precipitant (termed the mother liquor) which is tailor-made to provide crystallisation of the particular protein concerned. Both protein and mother liquor are applied to and mixed upon a glass cover slip which has been pre-coated using dimethyldichlorosilane solution to prevent the drop spreading out over the surface. The cover slip is then placed face down over a well containing approximately 1ml of mother liquor solution. A seal is formed

between cover slip and well using vacuum grease or oil added around the circumference of the well. This seal is reversible, allowing easy access to crystals in the drop once they have developed. The hanging drop is then left in its enclosed environment over a period of days or weeks, during which time a state of equilibrium is set up within the well, where water diffuses between the diluted protein-mother liquor mix in the drop and the stock mother liquor present in the well. It is during this process of equilibration that the crystals form as a result of controlled protein precipitation within the drop. Appendix 1 summarises the events occurring within the drop during equilibration which lead to crystal formation.

3.2.5 Crystal soaking, freezing and data collection

Cryoprotectant solution comprising of 15% PEG-4000, 20mM sodium phosphate, pH7.4, 10mg/ml reduced β-nicotinamide adenine dinucleotide (NADH, a ligand for momordin) and 40% glycerol was gently added to momordin crystals to a final glycerol concentration of greater than 30%. Crystals were left soaking in ligand-cryoprotectant solution for one hour and were then flash-frozen in an Oxford Cryosystems liquid nitrogen cryostream. Data collection took place on an Enraf-Nonius CuKα rotating anode X-ray source using a MAR Research image plate detector. 134 images were collected in total, where the oscillation angle was 1° per image, crystal-to-detector distance was 150mm, exposure time was 30 minutes per image and X-ray wavelength used was 1.5418Å. Spot intensities were collected and corrected using the data processing packages DENZO and SCALEPACK, and from there, processing of the β-NADH-soaked momordin data proceeded with the CCP4 (Collaborative Computing Project number 4) [167] programs SCALEPACK2MTZ, CAD, TRUNCATE, FREERFLAG, SFALL, SIGMAA, FFT and EXTEND, where phase information was provided from the native momordin structure [127] whose coordinates were obtained from the Protein Data Bank database [37]. Refinement was carried out initially with the CCP4 program RESTRAIN

and later with X-PLOR [170], and all graphical visualisation of three-dimensional structure was performed using QUANTA [168]. Validation of the final structure was achieved using the CCP4 program PROCHECK, which searches the structure for incorrect bond lengths, angles, mis-oriented side chains, etc. so that these can be corrected in the model.

Details of the X-ray crystallography programs listed above are provided in section 2.3.2.

It should be noted that coordinates for the ligand nicotinamide, as used in the model, were obtained from the chemical structural database [184].

3.3 MhpC

Growth Media/buffers

New Minimal Medium [185] – Basic Mix:

7.5mM Ammonium sulphate
8.5mM Sodium chloride
55mM KH_2PO_4
100mM K_2HPO_4
1mM Magnesium sulphate
20mM D-Glucose
1mg/L Ca^{2+} ions
1mg/L Fe^{2+} ions
1μg/L Cu^{2+} ions
1μg/L Mn^{2+} ions
1μg/L Zn^{2+} ions
1μg/L MoO_4^{2-} ions
10mg/L Thiamine
10mg/L Biotin
50mg/L 18 L-amino acid mix (minus L-methionine and L-cysteine) in 137mM NaCl, 2.5mM KCl, 10mM Na_2HPO_4 and 1.76mM KH_2PO_4, pH 7.0
100μg/ml Ampicillin

Luria Broth (LB) – per litre:

10g Tryptone
5g Yeast extract
5g Sodium chloride
1ml 1M Sodium hydroxide

LB plates – per litre:

10g Tryptone
5g Yeast extract
5g Sodium chloride
1ml 1M Sodium hydroxide
15g Agar

2X Sample buffer:

20mM Tris, pH8.0
2mM EDTA
5% Sodium dodecyl sulphate
10% β-mercaptoethanol
0.02% Bromophenol blue

N.B. Cautionary note: Selenomethionine is used repeatedly during this work. It is highly toxic if inhaled or exposed to skin, thus protective clothing, including gloves, filter mask and goggles must be worn when handling.

3.3.1 Transformation of the pIPC plasmid into the methionine auxotroph *E. coli* B834 (DE3) pLysS

The *mhpC* gene was ligated into the pIPC vector, where expression was under the control of the *lac* promoter and selection was ampicillin-dependent. The plasmid was constructed and digestion performed, by José L. Garcia , Consejo Superior de Investigaciones Científicas, Madrid, Spain. Transformation of the MhpC construct into the methionine auxotroph, *E. coli* B834 (DE3) pLysS (Novagen, catalogue number 69289-3) was achieved using the following method:

10-20ng plasmid was added to a pre-chilled, sterile 1.5ml eppendorf and placed on ice. Competent cells were rapidly thawed by hand and 100µl added to the DNA. After a brief mixing, via tube-flicking, the sample was placed on ice for 10 minutes. The cells were then heat-shocked via immersion in a water bath at 42°C for 2 minutes, immediately followed by a return to the ice for a further 10 minutes. 1ml of Luria Broth (LB) was next added to the tube which was then incubated at 37°C on a shaker set at 250rpm. The cells were next spun down in a bench microfuge at 13000 rpm for 1 minute and resuspended in 200µl LB to concentrate the sample. Various quantities of these stock cells (5µl, 20µl and 100µl) were then spread onto LB-ampicillin (100µg/ml) agar plates and these were incubated at 37°C for 12-16 hours.

3.3.2 Trial expression of MhpC in LB medium

A trial was set up to confirm that expression of MhpC was occurring before incorporating expensive selenium into the growth medium: An overnight culture was grown up at 37°C, 250rpm from a picked colony in 10ml LB-ampicillin (100µg/ml) medium. 1ml of the culture was then transferred to 800ml LB-ampicillin (100µg/ml), and cells grew up at 37°C, 250rpm.

After six hours of growth, the cell optical density at 600nm was 0.6-0.7, at which point the cells were induced with the addition of fresh, filter-sterile isopropyl-β-D-galactopyranoside (IPTG) to a final concentration of 0.6mM. Cells were then left to grow for 4 hours at 37°C and 7ml samples were collected before, and at time= 1, 2, 3 and 4 hours after induction. These were then spun down and mixed with 100µl 2X SDS sample buffer for SDS-PAGE.

3.3.3 Optimisation of LB concentration in growth media for selenomethionine MhpC preparation

Selenomethionine is naturally toxic to cells, and so in order to incorporate it into proteins, initial cell growth must be given a kick-start. Thus, very low levels of LB **containing L-methionine** are introduced into growth media, so that initial cell growth is favoured due to the incorporation of native L-methionine into protein rather than the potentially toxic selenomethionine. It is important to optimise these levels of LB in the growth medium used to prepare the selenomethionine MhpC protein so that just the right amount of L-methionine is added to drive translation, but not too much so that the selenomethionine is not introduced into the protein (L-methionine out-competes selenomethionine for incorporation into protein). The optimisation of the LB (and thus L-methionine) concentration to that value which allows ubiquitous incorporation of selenomethionine into protein during cell growth, is potentially an expensive procedure, due to the high cost of selenomethionine. Thus, optimisation of the LB (and hence L-methionine) concentration in this work was performed on a small scale, using minimal selenomethionine. The methionine-deficient growth medium used in this work is called New Minimal Medium (NMM) (see Growth Media/Buffers section above).

Several small-scale growth media trials (A-L, plus control) were set up as shown below.

NMM variant	Basic NMM mix	+ 0.5g/L stock L-cysteine	+30mM stock DL-selenomethionine	+ water	100% LB added
A	3.11ml	1ml	0.1ml	5.74ml	-
B	3.11ml	1ml	0.1ml	4.79ml	0.95ml
C	3.11ml	1ml	0.1ml	3.79ml	1.95ml
D	3.11ml	1ml	-	5.84ml	-
E	3.11ml	1ml	-	4.89ml	0.95ml
F	3.11ml	1ml	-	3.89ml	1.95ml
G	3.11ml	-	0.1ml	6.74ml	-
H	3.11ml	-	0.1ml	5.79ml	0.95ml
I	3.11ml	-	0.1ml	4.79ml	1.95ml
J	3.11ml	-	-	6.84ml	-
K	3.11ml	-	-	5.89ml	0.95ml
L	3.11ml	-	-	4.89ml	1.95ml
Control	-	-	-	-	9.95ml

DL-Selenomethionine was prepared commercially by Sigma (S-3875).

These growth media trials were designed to find the optimum growth medium which could be used in a large scale protein preparation to incorporate high levels of selenomethionine into proteins. The effect of various concentrations of L-methionine (added to the mix via LB) was examined in the presence and absence of DL-selenomethionine in order to determine a limiting concentration of LB which could be used to kick-start cell growth, but which would then become depleted, allowing 'exclusive' incorporation of selenomethionine into the cellular proteins. The necessity of L-cysteine in the growth medium was also examined since a personal correspondence suggested it might be a source of L-methionine, which could subsequently compete with the selenomethionine for incorporation into protein.

To each 9.95ml of growth medium, 50µl of overnight cells grown in 100% LB was added. This addition made each media volume up to 10ml, and it also adjusted the LB levels in each reaction to either 0.5%, 10% or 20%. These samples were next incubated at 37°C, in a shaker set at 250rpm. Samples were induced with IPTG to a final concentration of 0.6mM when their optical density at 600nm reached 0.6-0.7, and cells were harvested 4 hours after

induction for most samples, except tubes A and G where cells were harvested 8 hours after induction and J and D, where no induction took place. Harvested cells were mixed 1:1 with SDS-PAGE sample buffer and MhpC was visualised using 8-25% polyacrylamide SDS-PAGE Phastgels.

3.3.4 Large-scale expression and purification of selenomethionine MhpC

A 10ml overnight culture of the *E. coli* B834 cells containing the MhpC construct was grown up in 100% LB medium at 37°C, and at a shaker speed of 250rpm. 4ml of this was then added to 796ml of NMM containing 0.3mM DL-selenomethionine. Cells were then grown up again at 37°C and 250rpm, and 11 hours after incubation, the optical density of the cells at 600nm was 0.6, at which point induction took place via a final IPTG concentration of 0.6mM. Cells were harvested 8 hours later by centrifuging at 7227xg and at 4°C for 30 minutes. Cell pellets were resuspended in degassed (to minimise selenium oxidation) lysis buffer, at pH 7.0, containing 50mM potassium phosphate, 0.2mM EDTA, 5mM fresh PMSF (pre-prepared in 100% methanol), and 10mM β-mercaptoethanol, and were sonicated as follows: 20seconds on, 1 minute off, x12 cycles. Cell debris was spun down at 7740xg for 30 minutes at 4°C. Partial-purification was achieved via the addition of 25% (from salting-out table) ammonium sulphate which was stirred into the protein solution at 4°C for 30 minutes. Precipitated protein, including MhpC was then separated from the soluble fraction by centrifugation at 23700xg for 30 minutes at 4°C, and the resultant pellet resuspended in degassed anion exchange buffer containing 10mM β-mercaptoethanol, 0.2mM EDTA and 50mM potassium phosphate, pH 7.0. Dialysis was then carried out overnight at 4°C in a sealed beaker, against degassed 5L anion exchange column buffer.

After dialysis, the sample was spun down at 23700xg for 30 minutes at 4°C to remove any precipitated protein and the supernatant fraction was loaded onto a pre-washed, pre-

equilibrated Q-Sepharose anion exchange column (as used previously by Drs. Tom Robertson and Tim Bugg, Chemistry department, Southampton University). Anion exchange was performed using degassed anion exchange buffer with a sodium chloride gradient of 0-0.5M (250ml no salt buffer and 250ml 0.5M sodium chloride buffer). The flow rate was 5ml/minute and 5.3ml fractions were collected. Samples from key column fractions along with samples from various stages throughout the expression-purification process were run on an 8-25% polyacrylamide Phastgel for analysis of the overall success of the MhpC expression-purification process. MhpC samples were finally buffer-exchanged to remove salt in an aqueous solution containing 10mM DTT, which was added to prevent selenomethionine oxidation, and the samples were then concentrated down in preparation for crystallisation trials.

3.3.5 Crystallisation of selenomethionine MhpC

Crystallisation of selenomethionine MhpC was achieved using the hanging drop-vapour diffusion method (see section 3.2.4 for details). Both Hampton Research and Molecular Dimensions Limited crystal screening kits I and II were used to identify candidates for the selenomethionine MhpC mother liquor. Table 4 below, shows the solutions from the Molecular Dimensions Limited structure screening kits which yielded crystalline structures with the MhpC material.

Kit Number	Solution Number	Ingredients	Appearance
I	23	0.2M calcium chloride dihydrate*, 0.1M sodium HEPES, pH7.5*, 28% (v/v) PEG 400*	Star-shaped crystalline clusters
I	31	0.1M sodium HEPES, pH7.5*, 10% (v/v) 2-propanol, 20% (w/v) PEG 4000	Very dense crystal clusters
I	41	0.4M ammonium dihydrogen phosphate	Dark crystal clusters
I	45	4M sodium formate	Crystalline clusters
II	25	10mM cobalt chloride, 0.1M 2-(N-morpholino)ethanesulphonic acid (MES), pH 6.5, 1.8M ammonium sulphate	Very tiny crystalline structures
II	38	0.1M calcium chloride*, 0.1M sodium acetate, pH4.6, 30% (v/v) PEG 400*	Lots of very small crystalline structures

Table 4. A list of well solutions giving rise to crystal-like structures in an initial screen with MhpC. Note the ingredients present in more than one set of conditions (marked with an asterisk). The well solutions come from the Molecular Dimensions Limited Structure Screens I and II for X-ray crystallography.

Suitable crystallisation media were refined from initial "hits" using either grid- and/or factorial screening. Grid screening involves progressive variation from well to well of, typically, two parameters such as precipitant concentration and pH. Factorial screening involves a more systematic approach to determine which of the ingredients found in the "hits" are essential for crystallisation. Precipitant concentration, pH, protein concentration and additives are varied, so as to produce a series of wells of more widely differing constituents (see figure 19 below).

(a) Grid Screen

pH

Low ⟶ High

Precipitant concentration						
High	A1	A2	A3	A4	A5	A6
	B1	B2	B3	B4	B5	B6
	C1	C2	C3	C4	C5	C6
Low	D1	D2	D3	D4	D5	D6

(b) Factorial Screen

Precipitant Concentration	pH: Low		High		Protein Concentration
High	A1	A2	A3	A4	Low
	B1	B2	B3	B4	High
Low	C1	C2	C3	C4	Low
	D1	D2	D3	D4	High
	Present	Absent	Present	Absent	

Additive

Figure 19. Grid screen versus factorial screen for a suitable mother liquor. (a) The grid screen shows the effect of small changes in mother liquor parameters whereas (b) the factorial screen examines a wide spread of parameters.

Three distinct sets of crystallisation conditions giving rise to three different crystal forms of selenomethionine MhpC were identified:

1. 0.2M ammonium phosphate, 10mM DTT, 47% 2-methyl-2,4-pentanediol (MPD), 0.1M Tris, pH 8.5; protein concentration: 16-20mg/ml; growth temperature: 17°C. Crystals were evident after 2 days.

2. 50mM calcium chloride, 10% PEG-2000, 0.5% PEG-400, 50mM DTT, 0.1M sodium citrate, pH 5.6; protein concentration: 50mg/ml; growth temperature: 17°C. Crystals were visible after one week.

3. 50mM calcium chloride, 9.5-10% PEG-2000, 10mM DTT, 25% ethylene glycol, 0.1M N-(2-hydroxyethyl)piperazine-N'-(2-ethanesulphonic acid) (HEPES), pH 7.7; including 0.3ml Dow Corning 200/1cS fluid (BDH, product number 63002 4N):liquid paraffin (in a 1:1 mixture) layered over the top of the well solution; protein concentration: 25mg/ml; growth temperature: 4°C. Crystals appeared after 12 hours.

From this list of mother liquors, the optimal choice for experimentation was the set of conditions labelled "3". This particular mother liquor was carefully identified as follows:

(a) Molecular Dimensions Limited Structure Screens I and II were used with purified MhpC (15mg/ml) to screen previously-found "hits", or crystallisation conditions for other proteins.

(b) A list of "hits", or conditions with potential, which gave rise to crystal-like structures with MhpC protein, was drafted (the details of these "hits" are shown in table 4 above).

(c) Trends were identified from the list of "hits". Amongst these were: 0.1-0.2M calcium chloride, 28-30% (v/v) PEG-400, and 0.1M sodium HEPES, pH7.5 (see table 4).

(d) Optimisation of (i) the common ingredients identified in step (c), (ii) protein concentration and (iii) temperature, through grid and factorial screens.

(e) Addition of 10mM DTT to help protect the MhpC selenomethionines from oxidation.

(f) Addition of a 1:1 mix of Dow Corning 200/1cS fluid and liquid paraffin to give slower crystal growth and larger crystals.

3.3.6 Mass spectrometry on selenomethionine MhpC crystals

Native MhpC (expressed and purified by Dr. Tom Robertson, Chemistry department, Southampton University), selenomethionine MhpC and selenomethionine MhpC crystals were all evaluated by mass spectrometry to check that the selenomethionine had been successfully integrated into the MhpC protein molecules. In each case, the protein concentration was adjusted to between 0.1 and 1.0mg/ml and protein was placed in 50% acetonitrile and 1% formic acid. (In the case of the crystals, these were washed in fresh mother liquor for 15 minutes to remove contaminants and then dissolved in the acetonitrile:formic acid mix). Cone voltage, HV lens voltage and capillary voltage were adjusted to give an optimal multiple-charged envelope. The latter was then subjected to manual transform and/or maximum entropy analysis. Mass analysis was carried out on a Fisons Instruments VG Quattro II electrospray mass spectrometer.

3.3.7 Data collection on selenomethionine MhpC crystals

Data were collected on crystals grown under conditions "3" in section 3.3.5 above. These crystals were twinned and had to be cleaved in two in order to produce usable diffraction data. Cleaved crystal fragments were flash-frozen in liquid nitrogen where the mother liquor's ethylene glycol and PEG-2000 served as cryoprotectants. Data collection took place on a wavelength-tuneable synchrotron radiation beamline (station BM14) with a Mar Research charge-coupled device (CCD) –based detector at the European Synchrotron Radiation Facility (ESRF) in Grenoble, France. Initially, a fluorescence scan of a crystal was performed in order to confirm the presence of selenium within the protein crystal, and, more importantly, to allow selection of the wavelengths which would be most suitable for collection of anomalous data from the crystal. The latter is important since the form and position of the absorption edge peak varies from structure to structure and is dependent on the chemical environment of the anomalous scatterer(s) in the protein. Three wavelengths were chosen for M.A.D. data collection, and these gave rise to the so-called f " max, f ' min and remote data sets. The f " and f ' terms represent components of the anomalous scattering, and these vary with the incident X-ray wavelength. The f " component is directly proportional to the aforementioned fluorescence spectrum and the f ' component can be determined from the f " values using the Kramers-Kronig transformation [186,175]. The wavelength giving rise to the f " max data set is found from the peak of the fluorescence scan, where the anomalous diffraction signal is greatest. The wavelength for the f ' min data set corresponds to the wavelength at which the f ' component of the anomalous scattering is at its most negative value. Finally, the remote data set is collected at a wavelength remote from the selenium absorption edge, typically at a lower wavelength, where the difference between f ' at this wavelength and f ' min is maximal. This is important in order to be able to utilise the strongest anomalous signal due to the f ' component of the anomalous scattering [175].

With each of the three MhpC data sets collected, diffraction was from the same crystal, hence removing problems associated with non-isomorphous data. Each data set consisted of 240 images of 0.5° crystal oscillation per image, the crystal-to-detector distance was constant at 150mm, as was the exposure time, at 1 minute per image. The f" max, f' min and remote data sets were collected at X-ray wavelengths of 0.978689Å, 0.979035Å and 0.8855Å, respectively. The STRATEGY option of MOSFLM was employed to select an oscillation range of images where maximum completeness of unique data could be collected in the shortest time period. Preliminary data processing on these three data sets was performed, using the CCP4 programs MOSFLM, SORTMTZ, SCALA, TRUNCATE and HKLVIEW, which allowed precise identification of the crystal space group. Subsequent processing was achieved using the programs highlighted in figure 16(b), in section 2.3.3. Details of all these programs can be found in the main body of the text in section 2.3.3.

Chapter 4

Results

4. Results:

4.1 *Der p*1 and asthma

*Der p*1 is a key allergen derived from the house dust mite, *Dermatophagoides pteronyssinus*. It plays a major role in the onset of asthma attacks linked to house dust. Once inhaled, it triggers the immune response leading to airway inflammation. This section describes the expression of recombinant *Der p*1 fusion protein.

4.1.1 Fusion protein expression

*Der p*1 fusion protein expression was achieved at 30°C. However, the vast bulk of fusion protein was routed to an insoluble cellular compartment within the cell (most likely the inclusion bodies) as shown in the SDS-PAGE gels displayed below (figure 20 (a) and (b)).

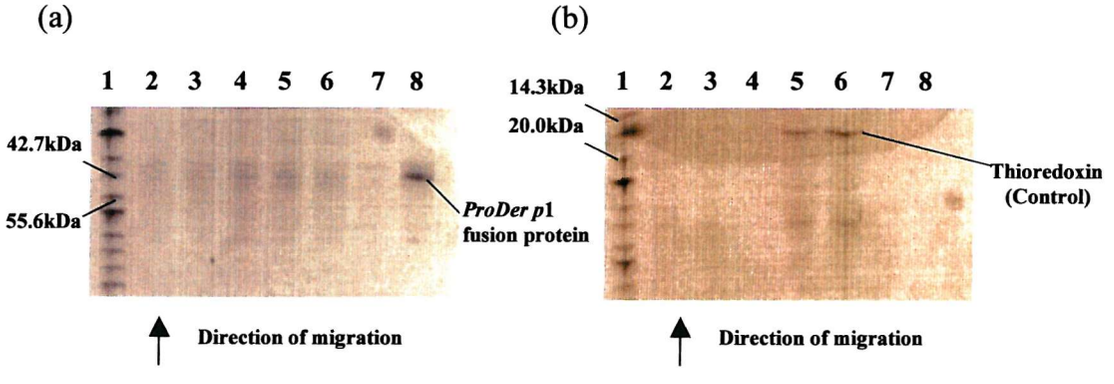


Figure 20. (a) 8-25% polyacrylamide SDS-PAGE Phastgel showing *proDer p*1-thioredoxin fusion protein expression at 30°C. Lane 1: Molecular weight markers (New England Biolabs, 2-212kDa, broad range), Lanes 2-6: *ProDer p*1-thioredoxin fusion protein expression in soluble fraction before, and at time = 1, 2, 3 and 4 hours after induction, respectively. Lanes 7 & 8: *ProDer p*1-thioredoxin fusion protein expression in insoluble fraction before, and 4 hours after induction, respectively. (b) 8-25% polyacrylamide SDS-PAGE Phastgel, showing thioredoxin expression with the same system under the same conditions (positive control). All lanes were labelled in the same way as with those shown in gel (a), except that the protein being expressed in gel (b) is a thioredoxin positive control.

From figures 20 (a) and (b), it can be seen that the molecular weights of the fusion protein and thioredoxin control, where expressed, were about 44-45kDa and 14-15kDa respectively. These values approximate the calculated molecular weight values of 47,253Da for the *proDer p*1 fusion protein and 15,117Da for the thioredoxin control.

4.1.2 Refolding of *proDer p*1

SDS-PAGE gels of *proDer p*1-thioredoxin fusion protein were run at two stages during the refolding process. The first displayed fusion protein present in samples immediately prior to the final dialysis step where excess guanidine is removed. This was simply to check for the presence of the fusion protein in the samples to ensure it had not been lost or degraded in-between its expression and its reaching this refolding stage (see figure 21 (a)). A second gel was then run at the end of the refolding procedure on samples which had been centrifuged at 13,000rpm for 15 minutes in a bench microfuge. It was postulated that any *proDer p*1-thioredoxin fusion protein present in the supernatant fraction must be soluble, refolded material (see figure 21 (b))

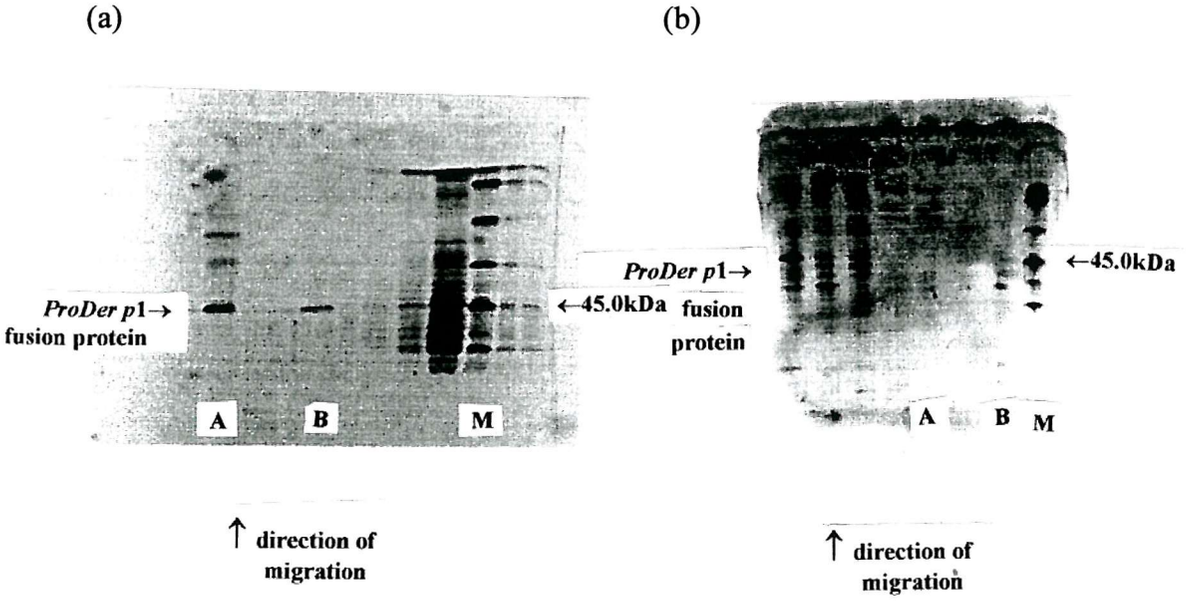


Figure 21. (a) 12% acrylamide SDS-PAGE gel of *proDer p*1 before dialysis to remove guanidine from the renaturation mix. M: Molecular weight markers (Bio Rad, low range). A and B are samples from two separate refolding mixes. (b) SDS-PAGE 8-25% acrylamide Phastgel containing samples from renaturation runs A and B. Lane M: Molecular weight markers. Lanes A and B: Supernatant samples after renaturation-centrifugation. 1µl sample was added per lane from a 5ml stock.

From gel (b), a very faint band is only just detectable in sample B at around 45kDa which is assumed to be soluble, renatured *proDer p*1 fusion protein. However, this band is very weak indeed. Sample C showed no bands at the 45kDa level, even with silver staining, suggesting the protein in that sample had not refolded.

4.1.3 Is *Der p*1 autocatalytic?

Following on from Taylor, *et al*'s work with the homologous cysteine protease papain, where it was suggested that papain is autocatalytic under certain conditions [183], the same experiment was reproduced with *Der p*1. Unfortunately, all gels run with the refolded *proDer p*1 sample in 'autocatalysis' buffer showed no mature *Der p*1 protein bands, even after silver staining.

4.2 Momordin and RIPs

Momordin is a toxin which acts as a ribosomal inhibitor. Here I describe its extraction from the seeds of the bitter gourd, *Momordica charantia*. The protein is purified in two steps as described by Husain, *et al* [127]. Purification data are shown below.

4.2.1 Gel Filtration of seed extract

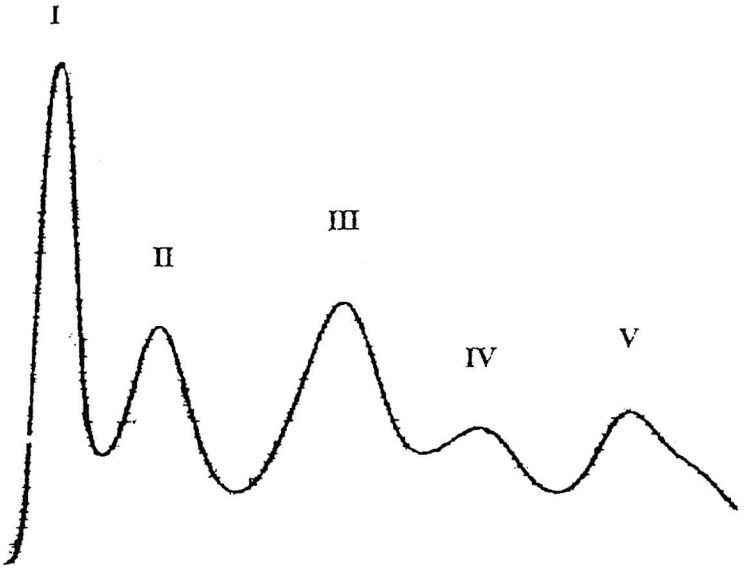


Figure 22(a) *Momordica charantia* seed extract samples I-V derived from the Sephadex G50 column. Momordin is present in peak II.

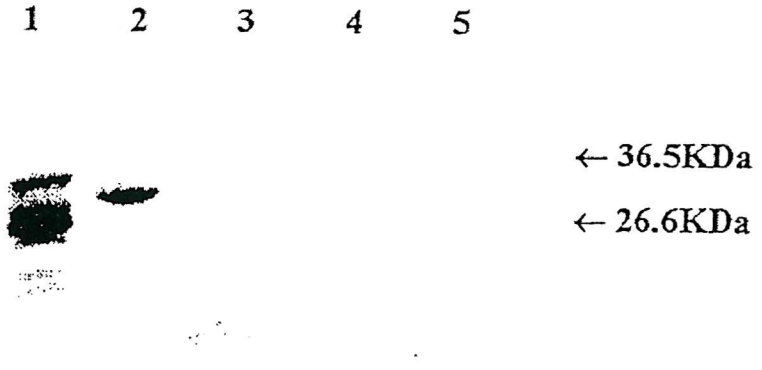


Figure 22(b) SDS-PAGE 8-25% Phastgel displaying samples 1-5 as derived from the Sephadex G50 column. Bands 1-5 correspond to peaks I-V shown in the trace above. Two molecular weight markers are displayed.

4.2.2 Cation exchange chromatographic separation of Peak II proteins derived from gel filtration

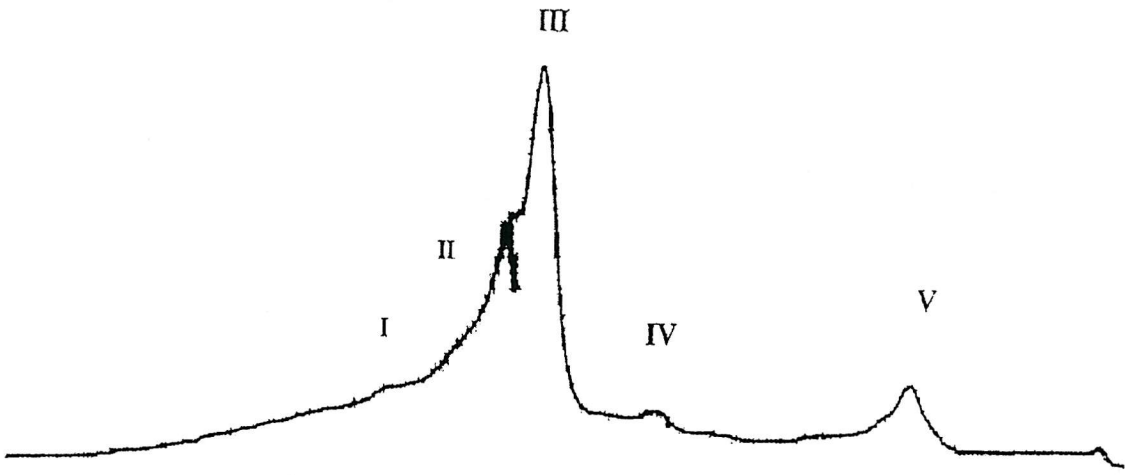


Figure 23(a) Elution profile displaying the various peaks (I-V) separated out using SP-Sepharose cation exchange chromatography. Momordin (α-momorcharin) is represented by peak III.

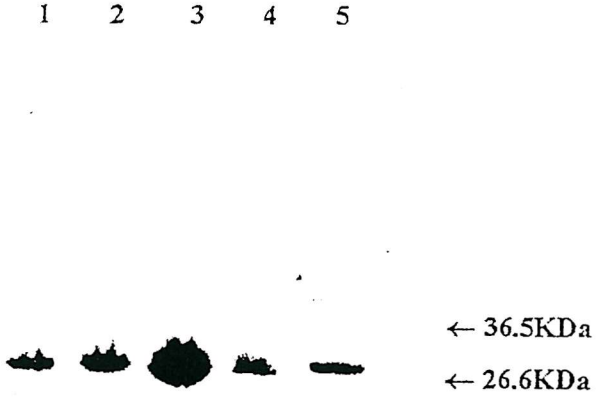


Figure 23(b) SDS-PAGE 8-25% Phastgel displaying fractions derived from the SP-Sepharose cation exchange chromatography run. Bands 1-5 represent the corresponding peaks from the above elution profile (see figure 23(a)). The positions of two molecular weight markers are shown.

Purification of momordin yielded about 50mg of protein from 100g initial seed sample.

4.2.3 Crystallisation of momordin

Crystallisation conditions for momordin were optimised from those published by Husain, *et al* [127]. Crystals appeared rhombohedral and were typically around 0.15mm^3 in size (see figure 24).

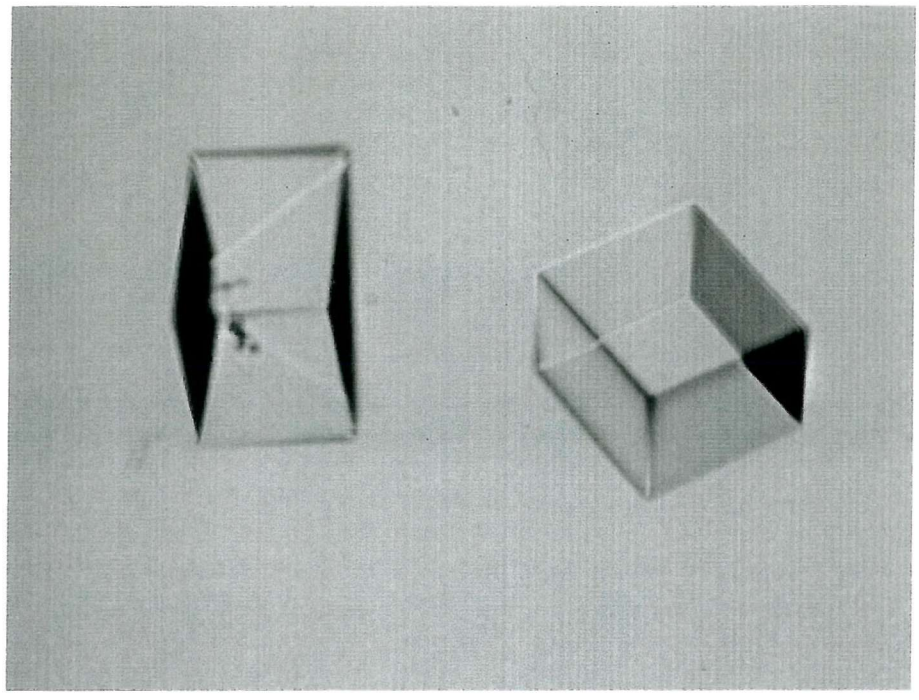


Figure 24. Momordin crystals (15mg/ml protein) grown in 6% PEG-4000 and 20mM sodium phosphate, pH 7.4. Crystal dimensions were 0.15mm^3. Note the rhombohedral architecture.

4.2.4 Data collection and processing

A typical diffraction pattern from the data set collected on the frozen momordin-βNADH-soaked crystal can be seen in figure 25 below. The diffracted spots could be resolved to 2Å resolution in-house.

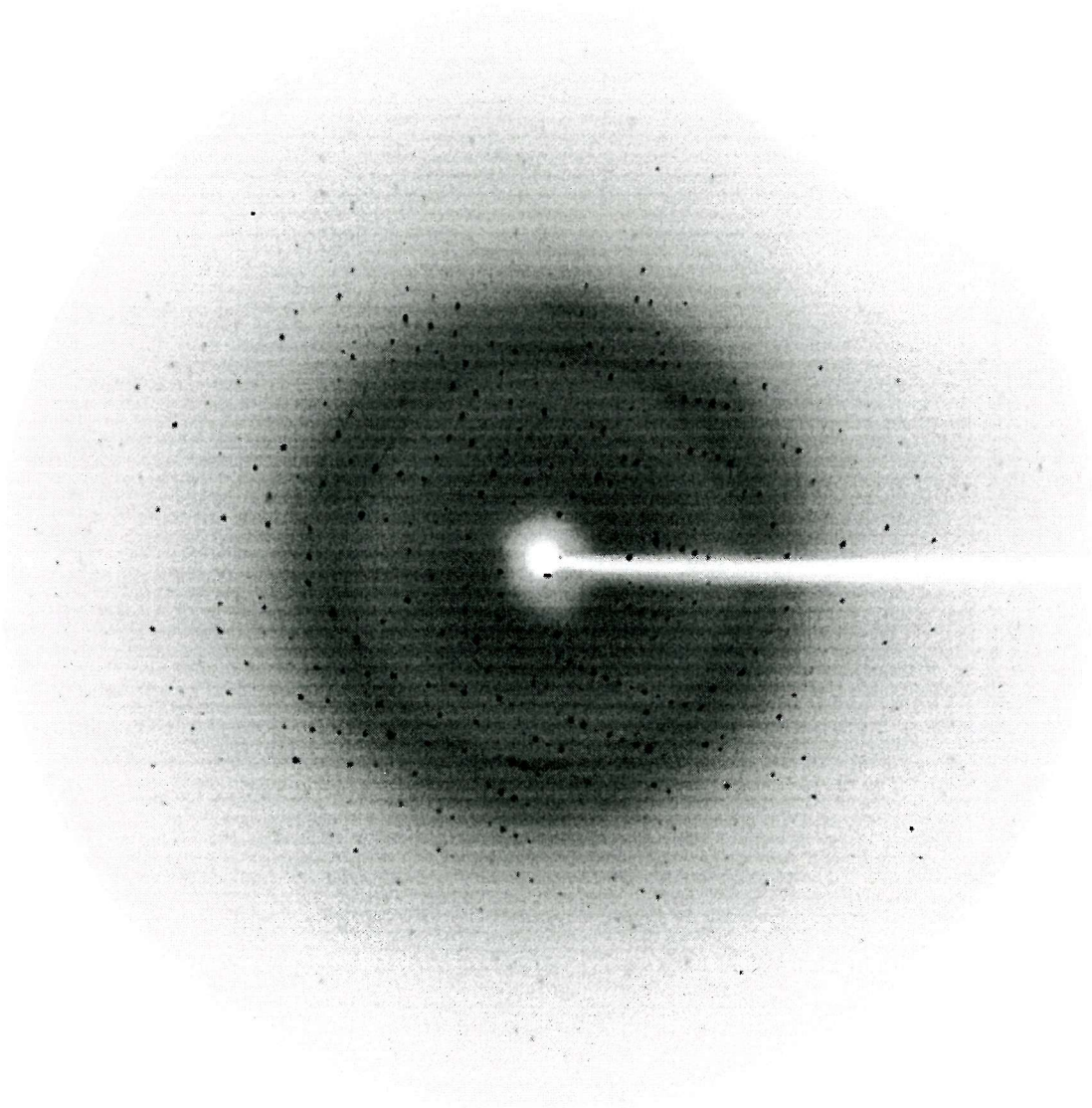


Figure 25. Diffraction pattern from the rhombohedral space group found in momordin-βNADH-soaked crystals.

Auto-indexing with DENZO ultimately yielded a rhombohedral, R3, space group (see table 5 below), and subsequent data processing gave refined cell parameters of:

$$a = b = 130.22\text{Å},\ c = 37.57\text{Å},\ \alpha = \beta = 90°,\ \gamma = 120°$$

Important statistical information relating to the success of the data processing is output from SCALEPACK (see table 6). Crucially, the overall completeness of the data is 96.7%, with reasonable intensity to background noise ratios for spots at up to 2Å resolution. Also important is the overall R_{merge} value for the data, which is 6.4%.

Lattice	Metric tensor distortion index	Best cell (symmetrized) Best cell (without symmetry restrains)					
primitive cubic	26.53%	75.88	76.20	76.22	62.50	117.34	62.77
		76.10	76.10	76.10	90.00	90.00	90.00
I centred cubic	20.22%	79.07	79.09	79.20	69.19	69.66	110.34
		79.12	79.12	79.12	90.00	90.00	90.00
F centred cubic	46.38%	129.83	79.20	90.33	52.02	89.99	89.73
		102.12	102.12	102.12	90.00	90.00	90.00
primitive rhombohedral	0.49%	76.20	75.88	76.22	117.34	117.50	117.23
		76.10	76.10	76.10	117.36	117.36	117.36
		130.06	130.06	37.51	90.00	90.00	120.00
primitive hexagonal	5.39%	75.88	76.22	37.51	80.51	80.78	117.34
		76.05	76.05	37.51	90.00	90.00	120.00
primitive tetragonal	18.31%	79.07	79.09	37.51	71.78	108.70	69.66
		79.08	79.08	37.51	90.00	90.00	90.00
I centred tetragonal	20.22%	79.07	79.09	79.20	69.19	69.66	110.34
		79.08	79.08	79.20	90.00	90.00	90.00
primitive orthorhombic	18.31%	37.51	79.07	79.09	110.34	108.22	108.70
		37.51	79.07	79.09	90.00	90.00	90.00
C centred orthorhombic	3.79%	79.09	129.92	37.51	89.82	71.78	89.71
		79.09	129.92	37.51	90.00	90.00	90.00
I centred orthorhombic	1.89%	37.51	75.88	130.30	89.94	89.91	80.78
		37.51	75.88	130.30	90.00	90.00	90.00
F centred orthorhombic	13.90%	79.07	89.87	130.30	89.91	90.01	127.75
		79.07	89.87	130.30	90.00	90.00	90.00
primitive monoclinic	5.06%	75.88	37.51	76.22	99.49	117.34	99.22
		75.88	37.51	76.22	90.00	117.34	90.00
C centred monoclinic	0.03%	79.07	130.30	37.51	90.09	108.70	90.01
		79.07	130.30	37.51	90.00	108.70	90.00
primitive triclinic	0.00%	37.51	75.88	76.20	117.23	99.65	99.22

Table 5. Distortion index table output from DENZO. The highlighted Bravais lattice, primitive rhombohedral, has a pretty good fit to the data (0.49%) and thus was selected for processing. (This was also the space group found for native momordin crystals [127]).

I/Sigma in resolution shells:

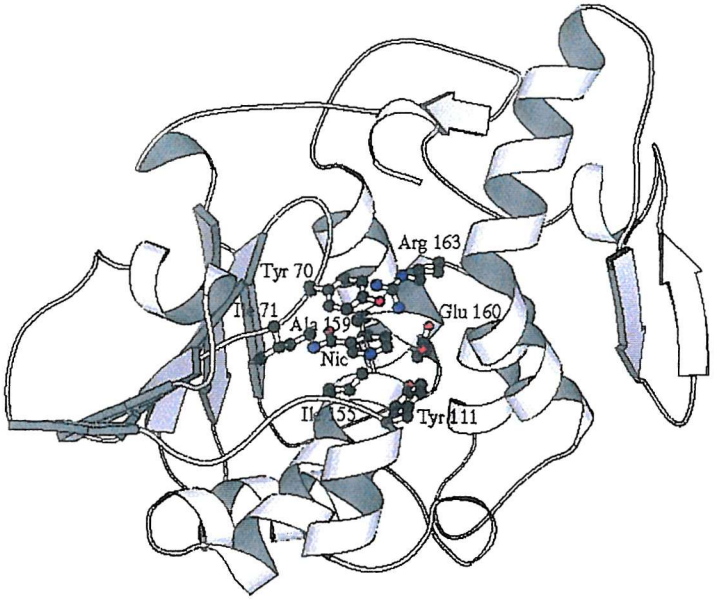
Lower limit	Upper limit	% of reflections with I / Sigma less than 0	1	2	3	5	10	20	+ >20	= total
40.00	4.31	0.6	1.6	1.6	1.9	2.5	5.3	19.0	78.7	97.8
4.31	3.42	0.1	0.4	0.8	1.1	1.7	6.4	21.7	76.8	98.5
3.42	2.99	0.4	1.2	2.5	4.0	6.7	16.2	38.7	60.0	98.6
2.99	2.71	1.0	2.3	4.4	6.6	11.6	23.8	50.7	48.1	98.8
2.71	2.52	1.8	4.1	7.2	10.7	18.0	34.6	62.3	36.6	98.9
2.52	2.37	1.6	4.1	8.5	12.6	23.1	42.3	70.8	28.3	99.1
2.37	2.25	2.3	6.9	13.0	19.0	30.8	50.7	76.7	22.1	98.8
2.25	2.15	4.6	10.8	18.0	25.3	37.7	57.0	81.6	16.9	98.5
2.15	2.07	6.1	14.6	24.1	32.8	47.6	68.9	90.0	8.8	98.9
2.07	2.00	8.6	18.0	26.7	34.9	46.7	62.6	76.4	2.4	78.8
All hkl		2.7	6.4	10.7	14.9	22.6	36.8	58.9	37.9	96.7

Table 6. Extract from SCALEPACK output file showing the completeness and strength of the momordin-βNADH-soaked data. The highlighted values are of particular note: 96.7% is the overall completeness of data from 40Å to 2Å resolution and (76.4%-34.9%) + 2.4% represents the data strength at 2Å resolution.

Yet more useful statistics are also output from TRUNCATE, which provides a list of unique spots from the collected data set which are converted to structure factor amplitudes, and which also provides a good estimate of the number of molecules in the asymmetric unit. For the momordin-βNADH-soaked data set, these values are 15,647 unique reflections, and one molecule per asymmetric unit.

Refinement of the momordin-βNADH structure yielded an R_{factor} of 24.0% and an R_{free} value of 26.1%, and the resultant electron density map was resolved to 2Å resolution. Using an Fo-Fc map, a region of electron density at the active site could be clearly identified which was not present in the native structure. Upon closer inspection, this region corresponded very well with a nicotinamide group, derived from the βNADH soaked into the momordin crystals (see figure 26).

(a)



MOMORDIN

(b)

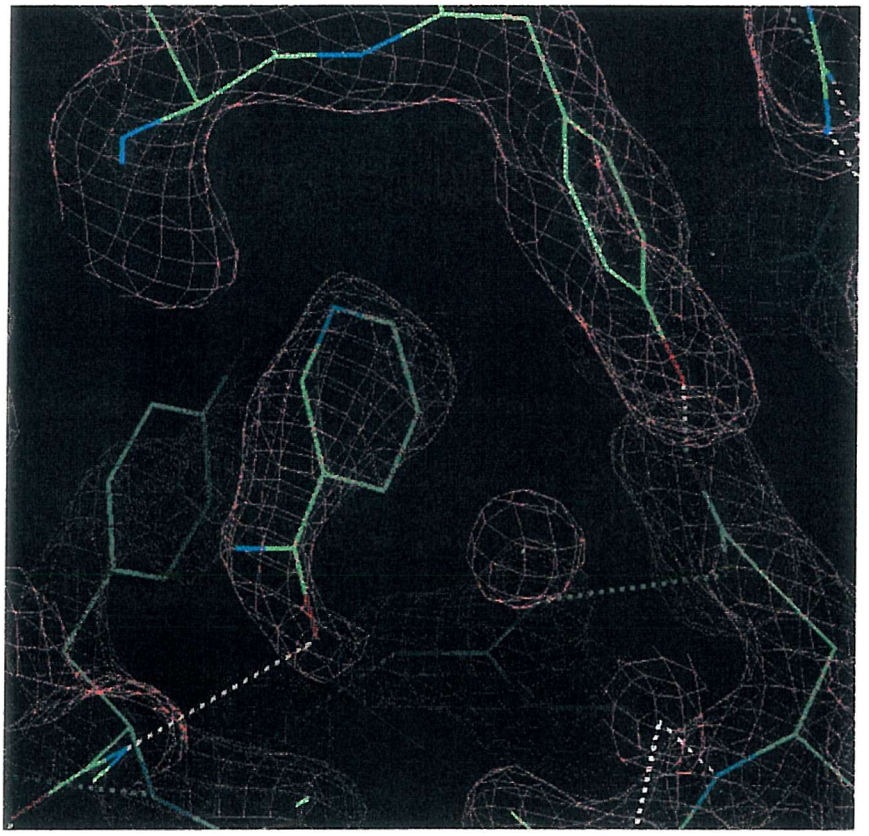


Figure 26. (a) MOLSCRIPT [35] representation of the momordin-βNADH structure with residues bordering the active site highlighted. "Nic" denotes the cleaved nicotinamide group. (b) Electron density map showing the presence of a strong, electron-dense region with nicotinamide modelled in at the active site of momordin. Hydrogen bonding is shown in the diagram. Notice how nicotinamide only forms a single hydrogen bond to the backbone of Ile-71. The nicotinamide group thus appears to be held in place principally by the hydrophobic Tyr-70 and Tyr-111. Picture captured from QUANTA [168].

A Ramachandran plot, created by the CCP4 structure validation program PROCHECK, for the momordin-nicotinamide model is shown below in figure 27.

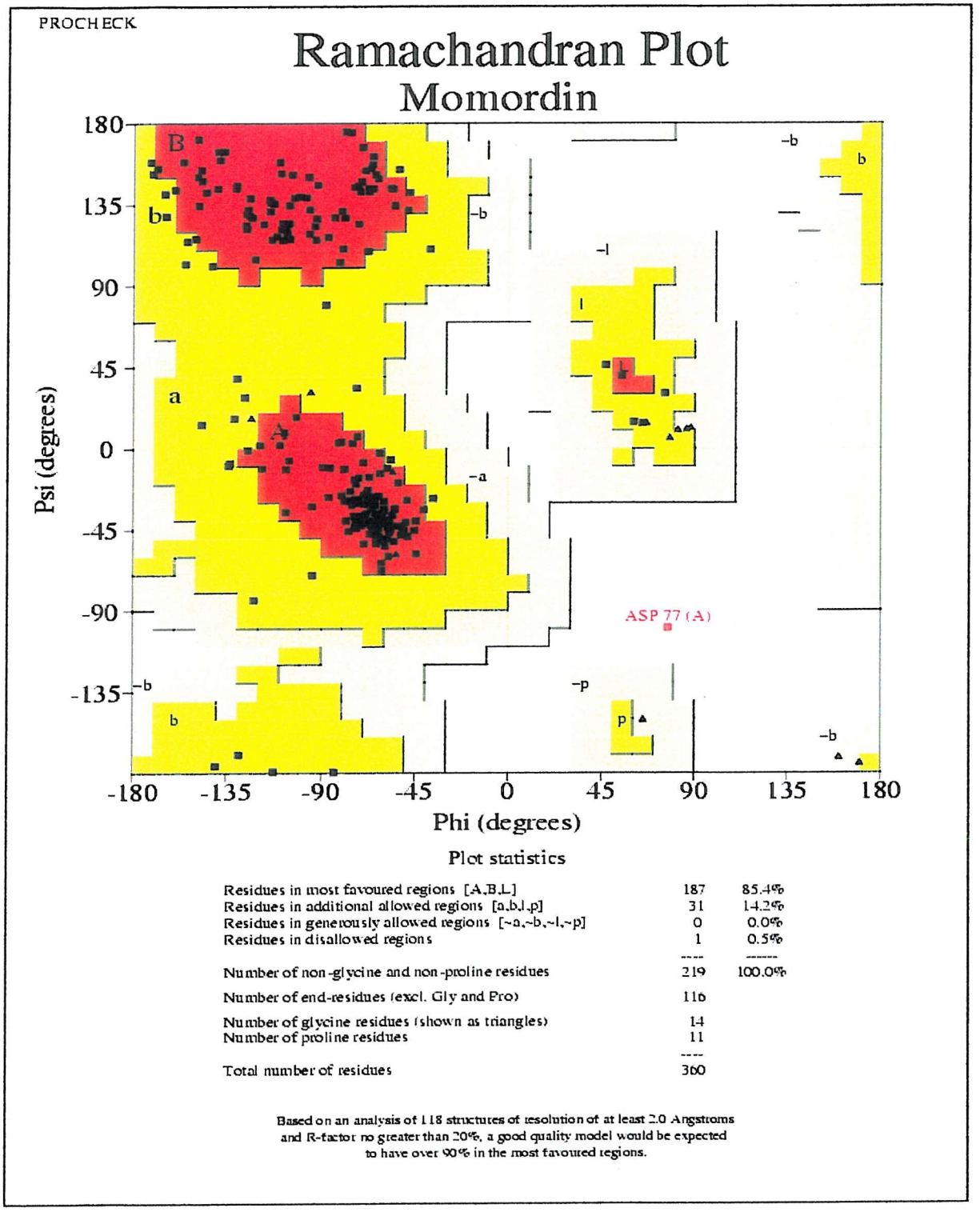


Figure 27. Ramachandran plot for the momordin-nicotinamide model. The aspartate 77 shown in the disallowed region of the plot is present in the structure at a very tight type II β-turn. It is also seen in this position in the native momordin structure, and similar occurrences have also been noted elsewhere ([187],[188]). This residue replaces glycine that commonly occurs here in homologous structures, such as ricin [127].

4.3 MhpC

MhpC is an enzyme of the *meta*-cleavage pathway for the degradation of 3-(3-hydroxyphenyl)propionic acids in *E. coli*. Its activity involves the breaking of a carbon-carbon bond, a rare catalytic event in nature.

4.3.1 Trial MhpC expression study on 100% LB

B834 (DE3) pLysS *E. coli* cells transformed with pIPC plasmid containing the *mhp*C gene grew up successfully on 100% LB, and strongly expressed protein with molecular weight corresponding to MhpC, upon induction (see figure 28 below).

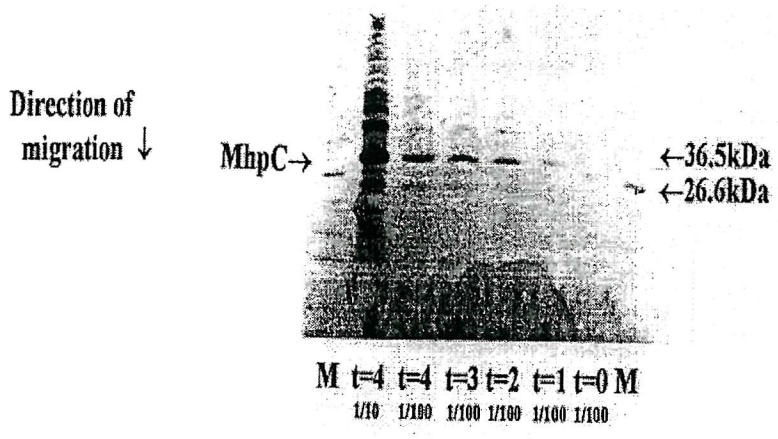


Figure 28. 8-25% SDS-PAGE polyacrylamide gradient Phastgel showing increasing levels of expression of MhpC with time. M: Molecular weight markers (New England Biolabs, broad range, catalogue number 7701S), t=0: before induction, t=1-4: time in hours after induction. 1/10 and 1/100 are dilutions of the culture with LB before adding to sample buffer. Note the progressive increase in MhpC concentration with time after induction.

4.3.2 Optimising the quantity of LB in the New Minimal Medium (NMM) growth mixture

Each of the small scale experiments, including mixtures of +/- cysteine, +/- selenomethionine and +0.5%, 10% and 20% LB, had a small sample removed between 5-5.5 hours after incubation began. These samples were analysed for cell density at 600nm wavelength (OD_{600}) at these times and the results are shown in table 7 for a comparison of the rapidity of growth in each case. The times at which an optical density at 600nm reached 0.6-0.7 are also shown, at which times induction took place.

Sample	Conditions	Time after incubation began when aliquot was removed	OD_{600} for aliquot	Induction time (after incubation began)
A	+Cys +Semet 0.5%LB	5 hours 26 minutes	0.089	11 hours 00 minutes
B	+Cys +Semet 10%LB	5 hours 19 minutes	0.230	7 hours 35 minutes
C	+Cys +Semet 20%LB	5 hours 10 minutes	0.329	6 hours 55 minutes
D	+Cys -Semet 0.5%LB	5 hours 25 minutes	0.100	-
E	+Cys -Semet 10%LB	5 hours 18 minutes	0.429	6 hours 07 minutes
F	+Cys -Semet 20%LB	5 hours 10 minutes	0.491	5 hours 55 minutes
G	-Cys +Semet 0.5%LB	5 hours 24 minutes	0.081	11 hours 40 minutes
H	-Cys +Semet 10%LB	5 hours 15 minutes	0.218	8 hours 05 minutes
I	-Cys +Semet 20%LB	5 hours 05 minutes	0.242	7 hours 05 minutes
J	-Cys -Semet 0.5%LB	5 hours 23 minutes	0.108	-
K	-Cys -Semet 10%LB	5 hours 15 minutes	0.440	6 hours 05 minutes
L	-Cys -Semet 20%LB	5 hours 00 minutes	0.424	6 hours 10 minutes
Control	100%LB	5 hours 00 minutes	0.631	5 hours 00 minutes

Table 7. Table showing the relative growth rates of each of the small scale trials involved in the optimisation of growth medium for the expression of selenomethionine MhpC. "Cys" denotes cysteine, whilst "Semet" denotes selenomethionine. Note that cells in samples D and J did not reach the inducible cell density of 0.6-0.7.

All samples eventually grew up to reach the inducible optical cell density, except D and J. Sample D had an OD_{600} of 0.110 12 hours after incubation began, and it had an OD_{600} of

0.095 7 hours later. No further growth took place. Similarly, sample J had an OD_{600} of 0.129 12 hours after incubation and a reading of 0.117 7 hours later.

All induced samples produced expression of a band corresponding to the molecular weight (about 31kDa) of MhpC which was comparable in density to that of the positive control. Protein expression continued for 4 hours after induction for all induced samples except the relatively slow growing samples A and G which grew up over 8 and 7.5 hours respectively, after induction, before samples were collected for running on SDS-PAGE gels.

Since media A and G (+Semet, 0.5% LB) gave rise to **inducible** cell growth, whilst media D and J (-Semet, 0.5% LB) gave only **limited** growth, then this showed that after the initial supplies of L-methionine in the minimal LB had been depleted, selenomethionine could subsequently be incorporated into bacterial proteins exclusively in place of the native L-methionine. At 10%/20% LB concentrations however, as can be seen from samples H and I (+Semet, 10%/20% LB) and K and L (-Semet, 10%/20% LB), which all grew up to inducible cell densities, L-methionine was heavily incorporated into the proteins in all these samples, thus dramatically reducing the quantity of selenomethionine incorporated. Therefore, a limiting concentration of LB at 0.5% in the initial growth medium was found to be optimal for maximum incorporation of selenomethionine into induced protein.

Samples D (+Cys) and J (-Cys) both used limited L-methionine (0.5% LB), and had no source of selenomethionine. Cells in both samples failed to grow to an inducible density, thus indicating that L-methionine could not be synthesised from L-cysteine in significant quantity in the *E. coli* methionine auxotroph, if at all.

4.3.3 Large scale expression-purification of selenomethionine MhpC with optimised growth medium

The SDS-PAGE gel below represents a summary of the overall expression-purification process with selenomethionine MhpC (see figure 29). Notice particularly how the purification is strongly dependent on the ammonium sulphate precipitation step.

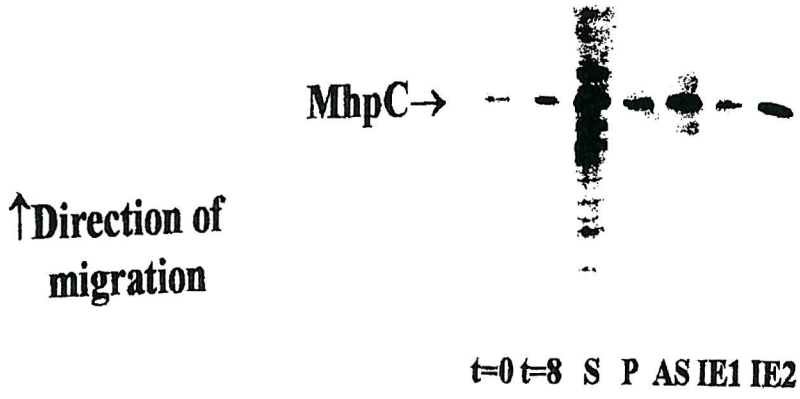


Figure 29. 8-25% polyacrylamide SDS-PAGE gradient Phastgel showing the presence of MhpC and other *E. coli* proteins at various stages of the expression-purification process. Lane descriptions are as follows: t = 0: negative control (before induction), t = 8: expression 8 hours after induction, S and P: Supernatant and pelleted fractions after cell lysis-centrifugation, A.S.: sample from aforementioned supernatant fraction, post-ammonium sulphate partial purification, centrifugation and dialysis. IE1 and IE2 samples from the peak fraction from Q-Sepharose anion exchange run.

4.3.4 Q-Sepharose anion exchange purification of MhpC

The trace below (figure 30 (a)) shows the elution profile from the Q-Sepharose anion exchange run. Two distinct major peaks are evident. Samples from each peak were run on an SDS-PAGE gel (see figure 30 (b) below).

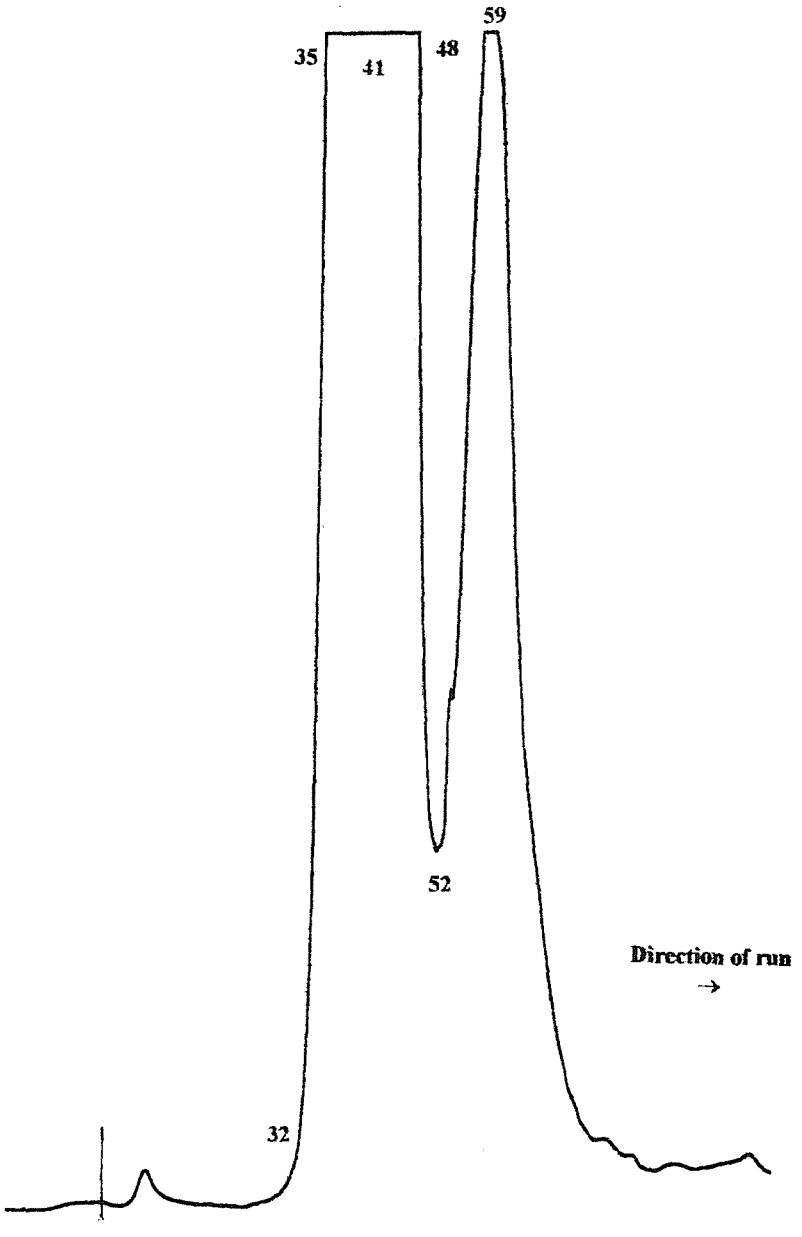


Figure 30 (a). Elution profile from Q-Sepharose column run with MhpC. Flow rate = 5ml/minute, chart speed = 1mm/minute, UV detector sensitivity at 280nm = 0.5, fraction size = 5.3ml. The trace was run from left to right. The first peak, due to its sheer size, is assumed to be the selenomethionine MhpC.

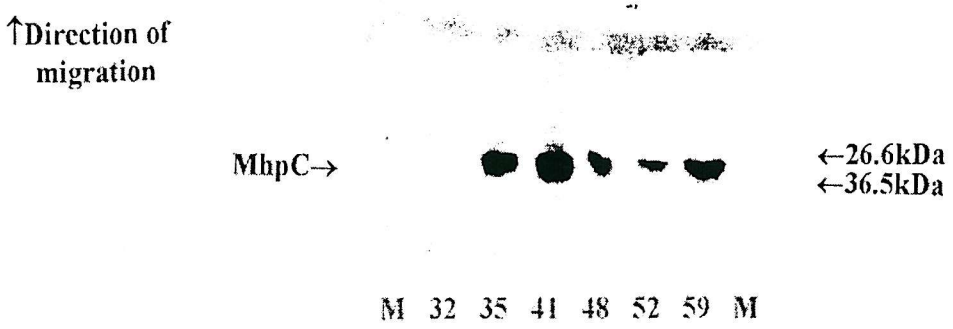


Figure 30. (b). 8-25% polyacrylamide SDS-PAGE gradient Phastgel showing select fractions from the peaks in the Q-Sepharose anion exchange trace shown in (a), above. M: molecular weight markers (New England Biolabs, broad range, code 7701S). All lane numbers on the gel correspond to fraction numbers from the anion exchange column trace. Samples were chosen from each of the two main peaks shown above, whose peak fractions were numbers 41 and 59. Note that the protein is of similar molecular weight in each peak and that 1μl of each 5.3ml fraction was added to the gel.

The overall yield of the major peak protein from the column was about 200mg. Since the first peak from the Q-Sepharose column is so large, it has been assumed to be a selenomethionine MhpC peak, and as such, it is this material which has been used for crystallisation trials, mass spectrometer analysis, etc.

4.3.5 Selenomethionine MhpC crystals

In section 3.3.5, three distinct crystal forms of MhpC were identified. Of these, number **1** mentioned there, was later found to be twinned and it was difficult to grow large enough crystals to allow chopping of these in half for data collection.

The crystal form number **2** was shaped like a rugby ball, where crystals were of reasonable size (about $0.2mm^3$). These showed no obvious twinning to the eye, and diffracted beautifully to 2Å resolution in-house. However, upon diffraction, these crystals gave excessively large unit cells, where one of the unit cell lengths was always many hundred Angstroms in length. As a consequence, data processing was unfortunately not feasible with these crystals. The third crystal form, number **3** in section 3.3.5, gave crystals which diffracted to around 2.9Å in-house, and which auto-indexed to a far smaller unit cell than with crystal form 2, and so these new crystals were far more workable. However, it was later shown that these crystals had twinning present within them. But even so, they grew sufficiently well (to about $0.2mm^3$) so that they could be cleaved into fragments which could then be used for data collection. So, all data collection presented in this chapter was performed on fragments of the crystal form 3. Intact crystals of the latter are shown in figure 31 below.

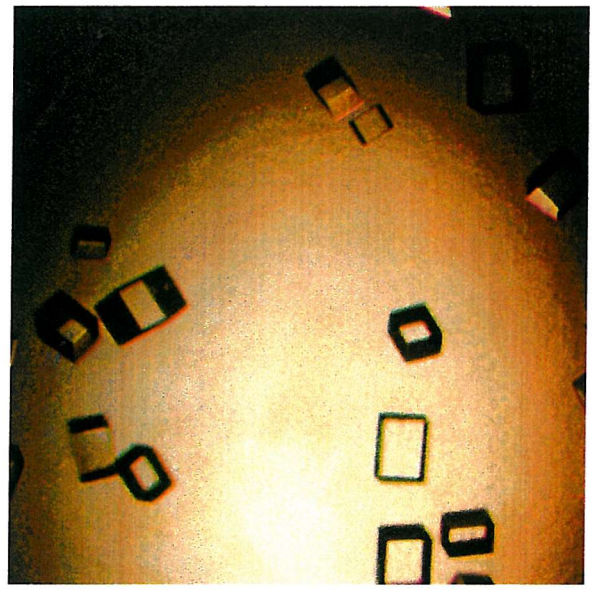


Figure 31. Selenomethionine MhpC crystals (approximately 0.2mm^3). Crystals were grown at 4°C by the hanging-drop method, in 50mM calcium chloride, 9.5-10% PEG-2000, 10mM DTT, 25% ethylene glycol, 0.1M HEPES, pH 7.7, with 0.3ml Dow Corning 200/1cS fluid:liquid paraffin (1:1) layered over the top of the well solution. Protein concentration used was 25mg/ml.

4.3.6 Mass spectrometry of MhpC

Mass spectra were obtained for native MhpC protein, selenomethionine MhpC protein and for selenomethionine MhpC protein crystals (see figures 32-24 below). Two major peaks identified from the traces have consistent molecular weights of about 31,830Da and 32,555Da which have been linked to the native MhpC and selenomethionine MhpC molecules. Selenomethionine MhpC (aqueous) protein and selenomethionine MhpC crystals show both of these peaks.

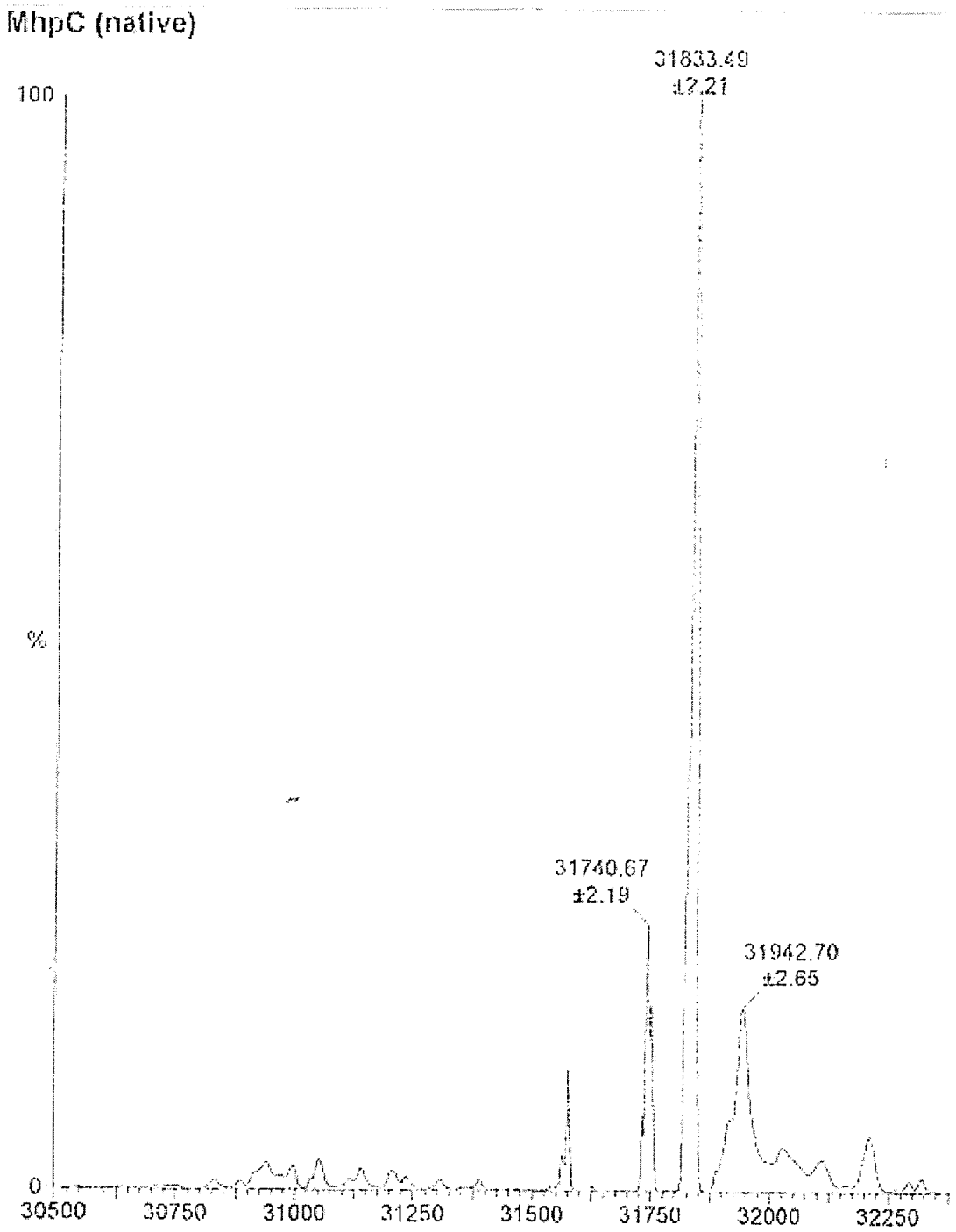


Figure 32. Mass spectrometer trace of native MhpC. Note the purity of the sample.

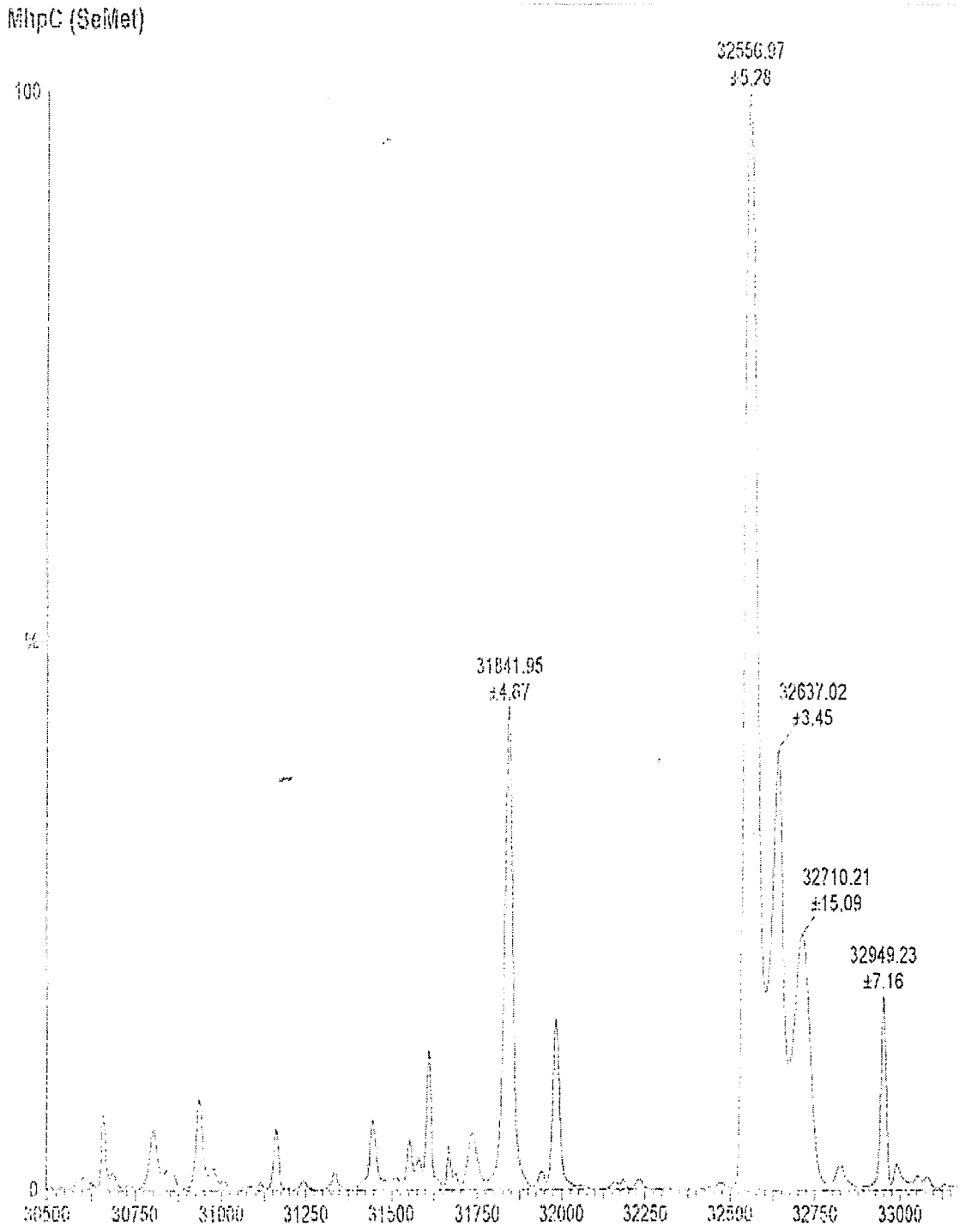


Figure 33. Mass spectrum of selenomethionine MhpC sample from the Q-Sepharose column fraction 41. Notice how large the 32,557Da peak is in comparison with the postulated native 31,842Da peak.

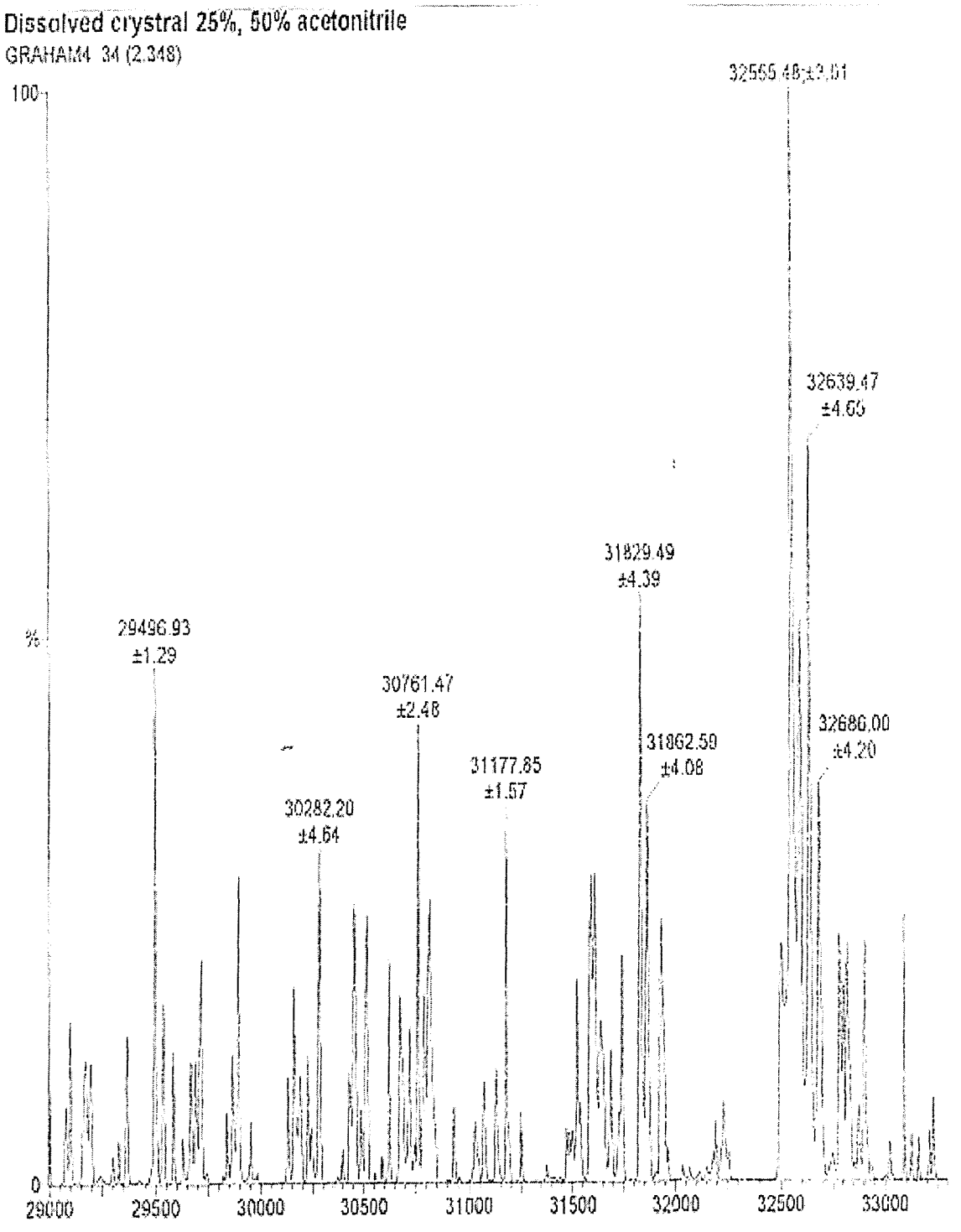


Figure 34. Mass spectrum of selenomethionine MhpC crystal. Once again, note the dominant 32,555Da peak and also the presence of the smaller 31,829Da native peak. The other peaks are of unknown origin, but may be modifications of the native and selenomethionine proteins.

Thus, selenium appears to be present in the MhpC protein.

4.3.7 Data collection on selenomethionine MhpC crystals

Figures 35 and 36 below show fluorescence scans of a selenomethionine MhpC crystal. Note the absorption edge in both figures which strongly indicates the presence of selenium in the crystals. The wavelengths used for data collection of the three data sets are found from these diagrams and correspond to (i) the f ' min amplitude, (ii) the f " max amplitude and (iii) to a remote wavelength away from the absorption edge.

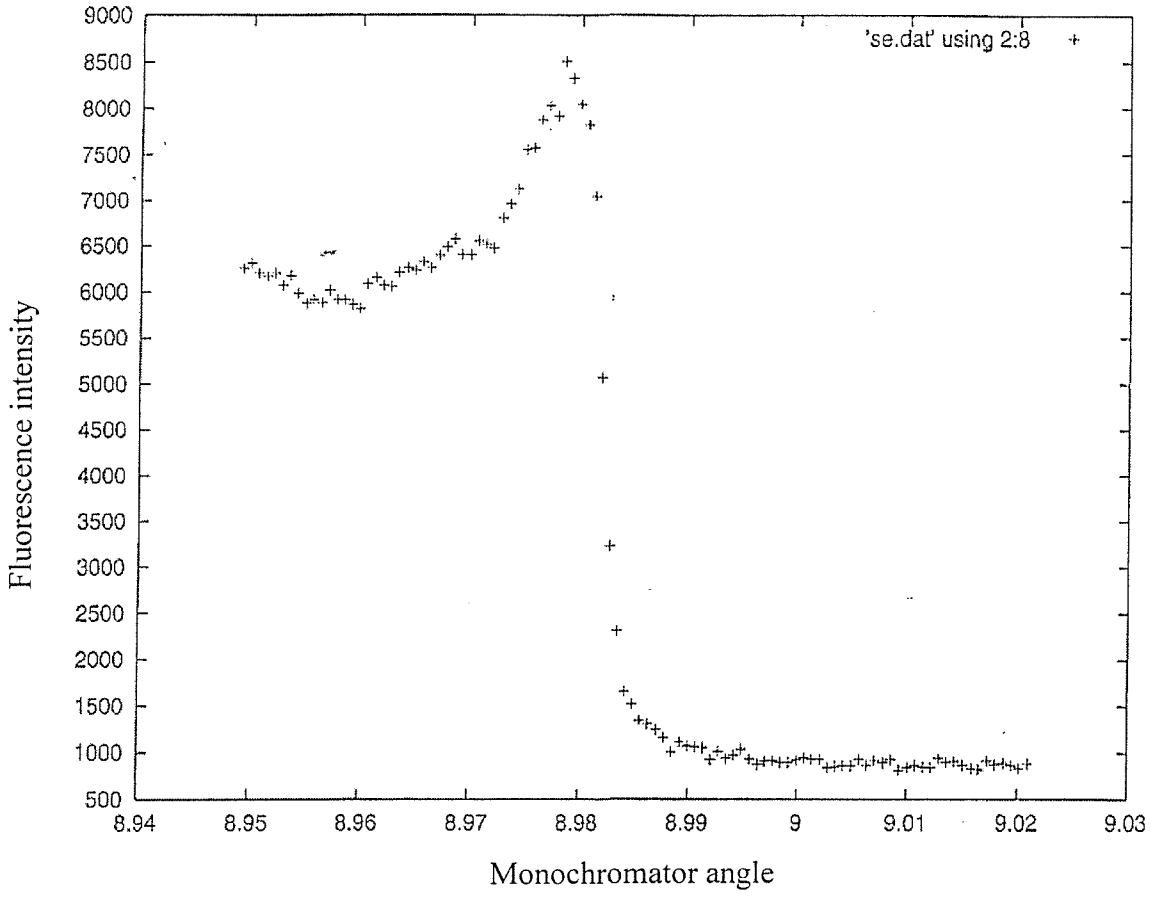


Figure 35. Fluorescence scan of selenomethionine MhpC crystal showing the absorption edge. The scale at the bottom is related to the wavelength of the X-rays used: the lower the value, the lower the wavelength.

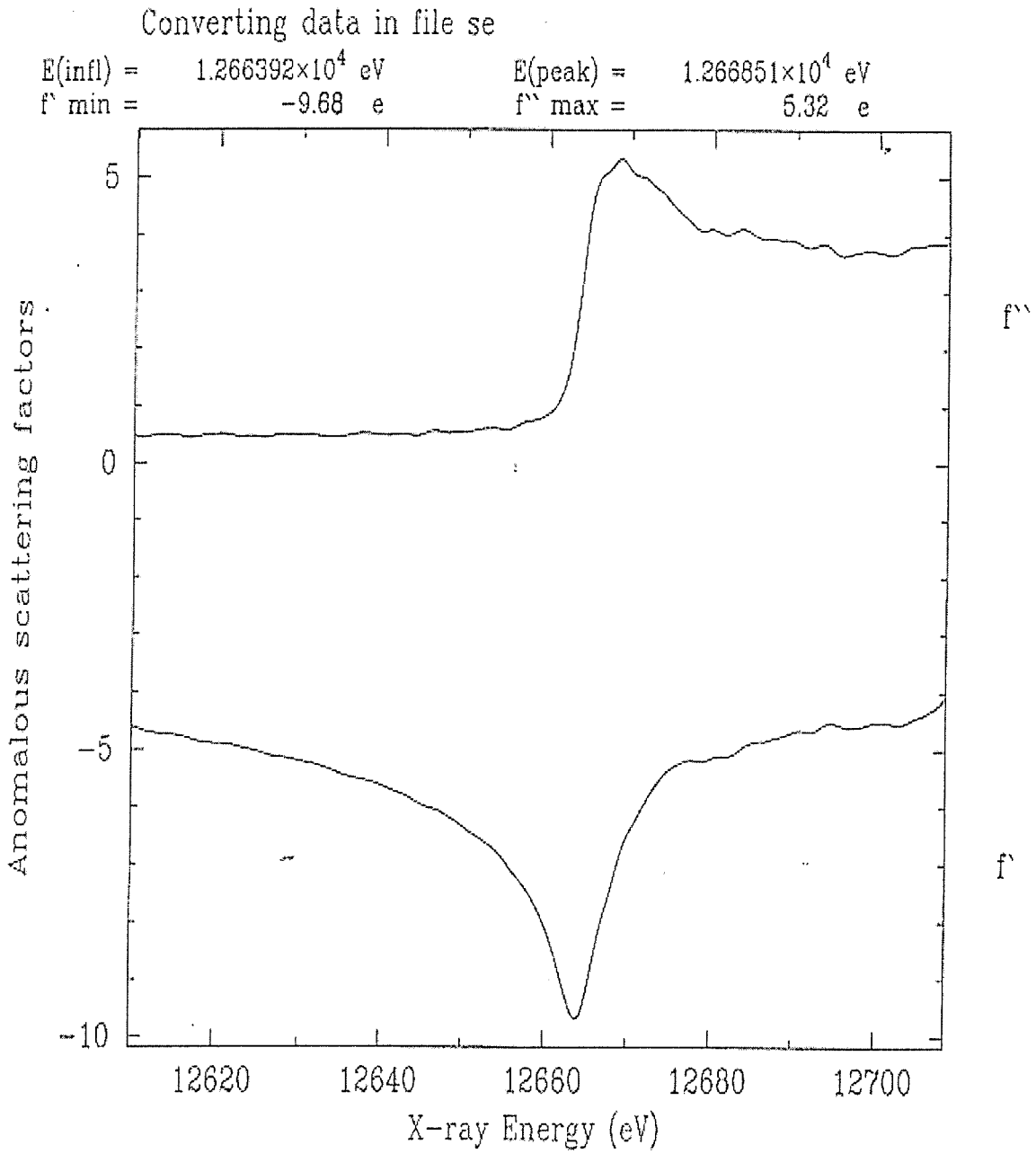


Figure 36. A split up version of the graph in figure 35. This shows the anomalous scattering factors f ' min and f " max amplitudes very clearly, where these are -9.68 electrons for f ' min and 5.32 electrons for f " max. As the X-ray energy increases along the bottom, so the wavelength decreases.

A typical diffraction pattern from one of the MAD data sets collected on the selenomethionine MhpC crystals is shown in figure 37. The spots recorded can be resolved to around 2.3Å resolution.

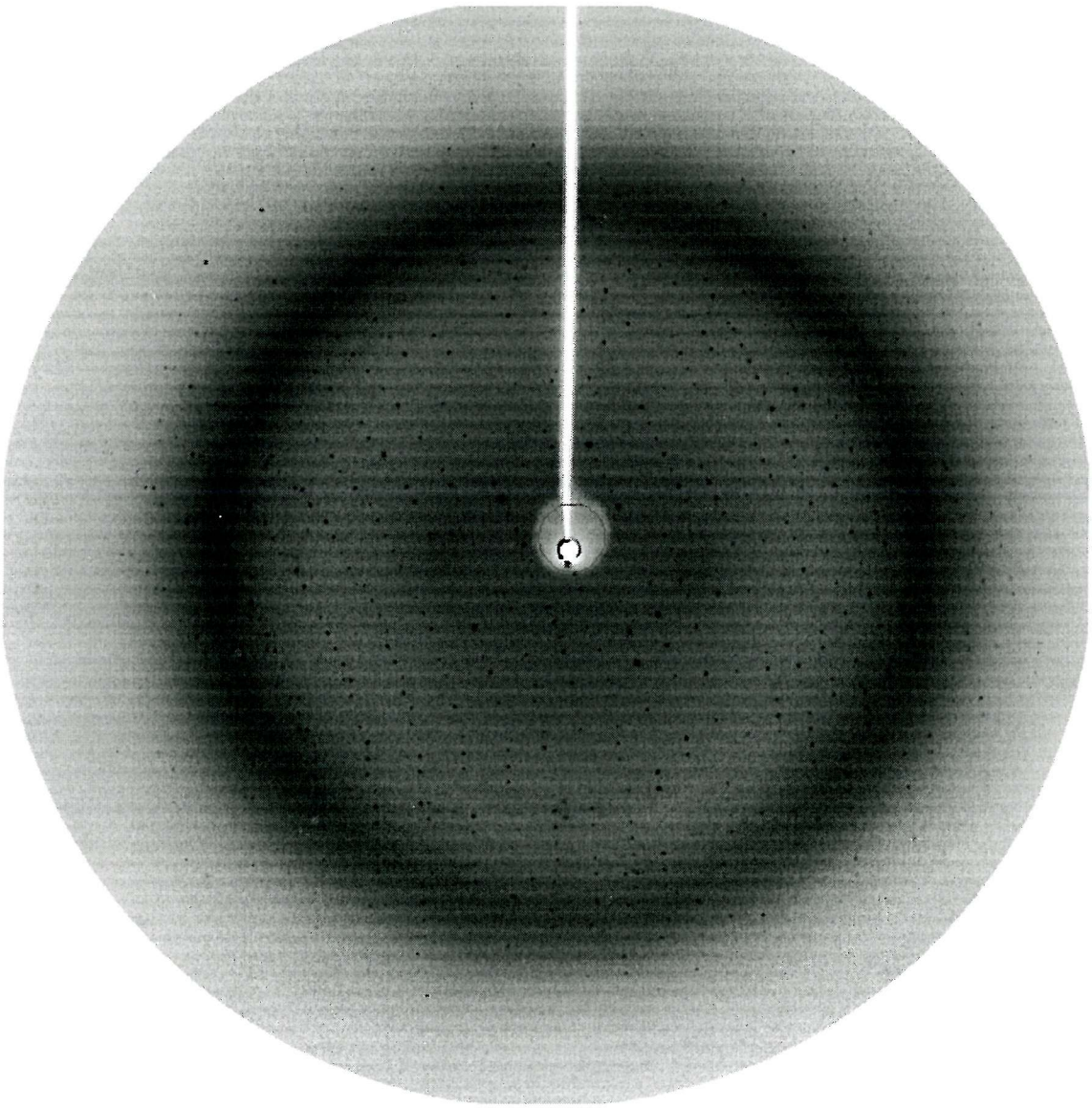


Figure 37. Diffraction image from selenomethionine MhpC crystal collected at the European Synchrotron Radiation Facility, Grenoble, France.

4.3.8 Initial data processing of selenomethionine MhpC data sets: the MOSFLM route

Auto-indexing with MOSFLM was ambiguous. Initially, the distortion index table seemed to suggest that the Bravais lattice was primitive orthorhombic for all three data sets (see table 8 below). However, later evidence from the programs SCALA and HKLVIEW showed that the lattice type was actually primitive monoclinic and that the space group was $P2_1$.

No	PENALTY	LATT	a	b	c	alpha	beta	gamma	Possible spacegroups
10	256	tP	70.99	70.99	157.85	90.0	90.0	90.0	P4,P41,P42,P43,P422, P4212,P4122,P41212, P4222,P42212 P4322,P43212
9	252	oC	100.68	101.55	157.85	90.0	90.0	90.0	C222,C2221
8	248	mC	101.55	100.68	157.85	90.0	90.2	90.0	C2
7	244	mC	101.55	100.68	157.85	90.0	90.2	90.0	C2
6	13	oP	62.41	79.56	157.85	90.0	90.0	90.0	P222,P2221,P21212,P212121
5	10	mP	62.41	79.56	157.85	90.0	89.8	90.0	P2,P21
4	9	mP	79.56	62.41	157.85	90.0	90.2	90.0	P2,P21
3	8	mP	62.41	157.85	79.56	90.0	90.5	90.0	P2,P21
2	2	aP	62.41	79.56	157.85	90.2	89.8	90.5	P1
1	0	aP	62.41	79.56	157.85	89.8	89.8	89.5	P1

Table 8. Extract from a distortion index table output from MOSFLM during the processing of the f ' min data set. Primitive orthorhombic (number 6 in the table) was the initial choice of Bravais lattice due to its high symmetry and low penalty. It was later seen from crystallographic symmetry and from identification of systematic absences, that the actual lattice for the crystal was primitive monoclinic (number 3 in the table), which was processed with b as the unique axis, in the $P2_1$ space group.

For each of the three data sets, the positional residuals upon data processing were good: f " max (e1): 0.038mm maximum, f ' min (e2): 0.036mm maximum and remote (e3): 0.041mm maximum.

Scaling yielded reasonable statistics. Of particular note were the I/sigma, % completeness and R_{sym} values (see tables 9-11).

TRUNCATE also gave some useful statistics: the number of unique reflections (also shown in the SCALA output), the post-refined cell dimensions, resolution range, space group and protein content of the unit cell, for each data set. Once again, this information is presented in tables 9-11, one for each of the three data sets.

N	1/d^2	Dmin(A)	Rfac	Rfull	Rcum	Ranom	Nanom	Av_I	SIGMA	I/sigma
1	0.0185	7.35	0.087	0.068	0.087	0.078	1789	13348.	2117.2	6.3
2	0.0370	5.20	0.079	0.067	0.083	0.074	3455	6740.	901.9	7.5
3	0.0555	4.25	0.083	0.067	0.083	0.069	4504	10700.	1562.3	6.8
4	0.0740	3.68	0.079	0.065	0.082	0.070	5372	8921.	1197.1	7.5
5	0.0925	3.29	0.081	0.066	0.081	0.073	6052	5860.	853.8	6.9
6	0.1109	3.00	0.081	0.067	0.081	0.077	6679	3311.	452.7	7.3
7	0.1294	2.78	0.095	0.082	0.082	0.094	7276	1543.	231.8	6.7
8	0.1479	2.60	0.124	0.108	0.084	0.129	7774	904.	203.7	4.4
9	0.1664	2.45	0.150	0.134	0.085	0.162	7171	647.	152.7	4.2
10	0.1849	2.33	0.159	0.136	0.085	0.232	710	643.	164.4	3.9
Overall:			0.085	0.070	0.085	0.077	50782	4387.	840.6	5.2
			Rfac	Rfull	Rcum	Ranom	Nanom	Av_I	SIGMA	I/sigma

N	Dmin	Nmeas	Nref	Ncent	%poss	Cm%poss	Mlplcty	AnomCmpl	AnomFrc	(Rsym)
1	7.35	5114	2023	92	96.3	96.3	2.5	89.2	92.1	0.087
2	5.20	9871	3809	141	99.8	98.5	2.6	94.0	94.1	0.079
3	4.25	12745	4883	128	99.4	98.9	2.6	94.1	94.6	0.083
4	3.68	15156	5794	137	99.7	99.2	2.6	94.7	94.8	0.079
5	3.29	17022	6514	131	99.0	99.2	2.6	93.9	94.7	0.081
6	3.00	18827	7177	126	98.8	99.1	2.6	93.6	94.5	0.081
7	2.78	20470	7831	132	99.2	99.1	2.6	93.8	94.3	0.095
8	2.60	21876	8347	127	98.5	99.0	2.6	93.2	94.4	0.124
9	2.45	21455	8904	125	98.8	99.0	2.4	80.7	81.6	0.150
10	2.33	5659	4132	50	43.4	90.8	1.4	7.5	17.2	0.159
Overall		148195	59414	1189	90.8	90.8	2.5	79.2	87.0	0.085
		Nmeas	Nref	Ncent	%poss	Cm%poss	Mlplcty	AnomCmpl	AnomFrc	(Rsym)

```
 * Number of Reflections =  59414

 * Cell Dimensions :

   63.032  155.278   79.425   90.000   89.519   90.000

 * Resolution Range :

   29.361 -     2.326 A

 * Space group = P21  (number     4)

 * Fraction of unit cell occupied by atoms =         0.464
```

Table 9. Key statistics output from SCALA (top 2 tables) and from TRUNCATE (asterisked parameters) for the e1 (f" max) data set, when processed with spacegroup $P2_1$. Values of particular note are the I/sigma at 2.33Å resolution, which is an indicator of spot strength there, the overall % completeness of the data, 90.8%, which shows whether there is sufficient data to allow processing to maximum resolution, the R_{sym}, which represents how close the values of symmetry-related spot intensities are to one another, and which is a marker of data quality, and then there is the multiplicity, or averaged number of symmetry-related spots per set in the collected data, which, where high, gives better final averaged intensity statistics for a given set of symmetry-related spots. TRUNCATE and SCALA both output the total number of unique spots collected, the final post-refined cell dimensions, the resolution range of the data, and the choice of space group. TRUNCATE alone calculates the unit cell solvent content, which is expressed as a percentage, and from which, the total protein content in the unit cell can also be determined and is also output.

N	1/d^2	Dmin(A)	Rfac	Rfull	Rcum	Ranom	Nanom	Av_I	SIGMA	I/sigma
1	0.0184	7.37	0.086	0.069	0.086	0.077	1776	13712.	2101.4	6.5
2	0.0368	5.21	0.078	0.067	0.082	0.074	3431	6876.	913.7	7.5
3	0.0552	4.26	0.082	0.066	0.082	0.070	4491	10813.	1569.6	6.9
4	0.0736	3.69	0.079	0.065	0.081	0.070	5326	9009.	1208.9	7.5
5	0.0920	3.30	0.081	0.066	0.081	0.073	6023	5975.	877.1	6.8
6	0.1104	3.01	0.081	0.067	0.081	0.076	6662	3378.	470.2	7.2
7	0.1288	2.79	0.096	0.083	0.082	0.093	7239	1555.	236.5	6.6
8	0.1472	2.61	0.122	0.105	0.083	0.127	7730	919.	209.0	4.4
9	0.1656	2.46	0.149	0.135	0.085	0.158	7294	642.	150.3	4.3
10	0.1840	2.33	0.162	0.138	0.085	0.229	750	643.	166.6	3.9
Overall:			0.085	0.070	0.085	0.077	50722	4439.	845.7	5.2
			Rfac	Rfull	Rcum	Ranom	Nanom	Av_I	SIGMA	I/sigma

N	Dmin	Nmeas	Nref	Ncent	%poss	Cm%poss	Mlplcty	AnomCmpl	AnomFrc	(Rsym)
1	7.37	5084	2015	94	96.3	96.3	2.5	89.1	92.0	0.086
2	5.21	9815	3783	136	99.7	98.5	2.6	93.8	94.0	0.078
3	4.26	12692	4868	129	99.6	99.0	2.6	94.4	94.6	0.082
4	3.69	15043	5754	138	99.5	99.2	2.6	94.4	94.7	0.079
5	3.30	16945	6482	129	99.1	99.1	2.6	93.9	94.6	0.081
6	3.01	18765	7156	127	99.0	99.1	2.6	93.9	94.6	0.081
7	2.79	20373	7793	133	99.2	99.1	2.6	93.8	94.3	0.096
8	2.61	21745	8310	125	98.6	99.0	2.6	93.1	94.3	0.122
9	2.46	21587	8844	125	98.6	99.0	2.4	82.5	83.5	0.149
10	2.33	5795	4191	50	44.3	91.0	1.4	8.0	17.9	0.162
Overall		147844	59196	1186	91.0	91.0	2.5	79.5	87.2	0.085
		Nmeas	Nref	Ncent	%poss	Cm%poss	Mlplcty	AnomCmpl	AnomFrc	(Rsym)

```
 * Number of Reflections =  59196

 * Cell Dimensions :

   63.071  155.362   79.477   90.000   89.562   90.000

 * Resolution Range :

   29.361 -     2.331 A

 * Space group = P21  (number     4)

 * Fraction of unit cell occupied by atoms =  0.463
```

Table 10. Key statistics output from SCALA and TRUNCATE for the e2 (f ' min) data set, when processed with spacegroup $P2_1$. (See table 9 legend for salient points to note).

N	1/d^2	Dmin(A)	Rfac	Rfull	Rcum	Ranom	Nanom	Av_I	SIGMA	I/sigma
1	0.0227	6.63	0.083	0.067	0.083	0.078	2392	17838.	2613.2	6.8
2	0.0454	4.69	0.083	0.074	0.083	0.070	4734	15409.	2369.4	6.5
3	0.0682	3.83	0.080	0.067	0.082	0.068	6164	19014.	2759.3	6.9
4	0.0909	3.32	0.080	0.067	0.081	0.068	7262	11818.	1703.8	6.9
5	0.1136	2.97	0.079	0.069	0.081	0.072	8296	6226.	863.0	7.2
6	0.1363	2.71	0.092	0.081	0.082	0.088	9138	2613.	385.5	6.8
7	0.1590	2.51	0.120	0.105	0.083	0.122	9883	1580.	388.4	4.1
8	0.1818	2.35	0.141	0.124	0.085	0.154	10566	1132.	238.5	4.7
9	0.2045	2.21	0.250	0.211	0.087	0.261	9351	742.	850.2	0.9
10	0.2272	2.10	0.221	0.200	0.087	0.389	123	605.	242.7	2.5
Overall:			0.087	0.075	0.087	0.078	67909	6248.	1360.4	4.6
			Rfac	Rfull	Rcum	Ranom	Nanom	Av_I	SIGMA	I/sigma

N	Dmin	Nmeas	Nref	Ncent	%poss	Cm%poss	Mlplcty	AnomCmpl	AnomFrc	(Rsym)
1	6.63	6667	2709	112	94.8	94.8	2.5	87.4	90.9	0.083
2	4.69	13454	5186	165	99.9	98.1	2.6	94.2	94.0	0.083
3	3.83	17368	6668	167	99.6	98.7	2.6	94.4	94.6	0.080
4	3.32	20461	7823	159	98.9	98.8	2.6	93.7	94.5	0.080
5	2.97	23337	8922	164	99.5	99.0	2.6	94.3	94.5	0.079
6	2.71	25690	9811	162	99.1	99.0	2.6	93.8	94.5	0.092
7	2.51	27745	10603	156	98.6	98.9	2.6	93.3	94.3	0.120
8	2.35	29754	11353	154	98.3	98.8	2.6	92.8	94.0	0.141
9	2.21	28447	12065	159	98.2	98.7	2.4	77.1	78.4	0.250
10	2.10	2422	2053	4	15.9	86.6	1.2	1.0	5.8	0.221
Overall		195345	77193	1402	86.6	86.6	2.5	77.5	89.3	0.087
		Nmeas	Nref	Ncent	%poss	Cm%poss	Mlplcty	AnomCmpl	AnomFrc	(Rsym)

```
 * Number of Reflections =  77193

 * Cell Dimensions :

   63.066  155.330   79.479   90.000   89.485   90.000

 *  Resolution Range :

    29.235 -     2.098 A

 * Space group = P21  (number     4)

 * Fraction of unit cell occupied by atoms =         0.463
```

Table 11. Key statistics output from SCALA and TRUNCATE for the e3 (remote) data set, when processed with spacegroup $P2_1$. (See table 9 legend for important parameters to note).

Final determination of the Bravais lattice and space group was achieved via information output from both SCALA and from HKLVIEW. Processing with P2, SCALA initially output intensity statistics which strongly indicated the presence of systematic absences along the reciprocal lattice k axis, and since the R_{merge} values output from SCALA were very poor (40-60%) when the data was processed with space group P222, but were far better when processed in $P2_1$ (see tables 9-11), this then suggested that $P2_1$ was the correct space group. This was confirmed by HKLVIEW, via pseudo-precession pictures which revealed the crystallographic symmetry pattern of a primitive monoclinic cell (see figure 38) and systematic absences along the reciprocal lattice unique k axis (when the data was processed in P2), which revealed the presence of a 2_1 screw axis in the data (see figure 39).

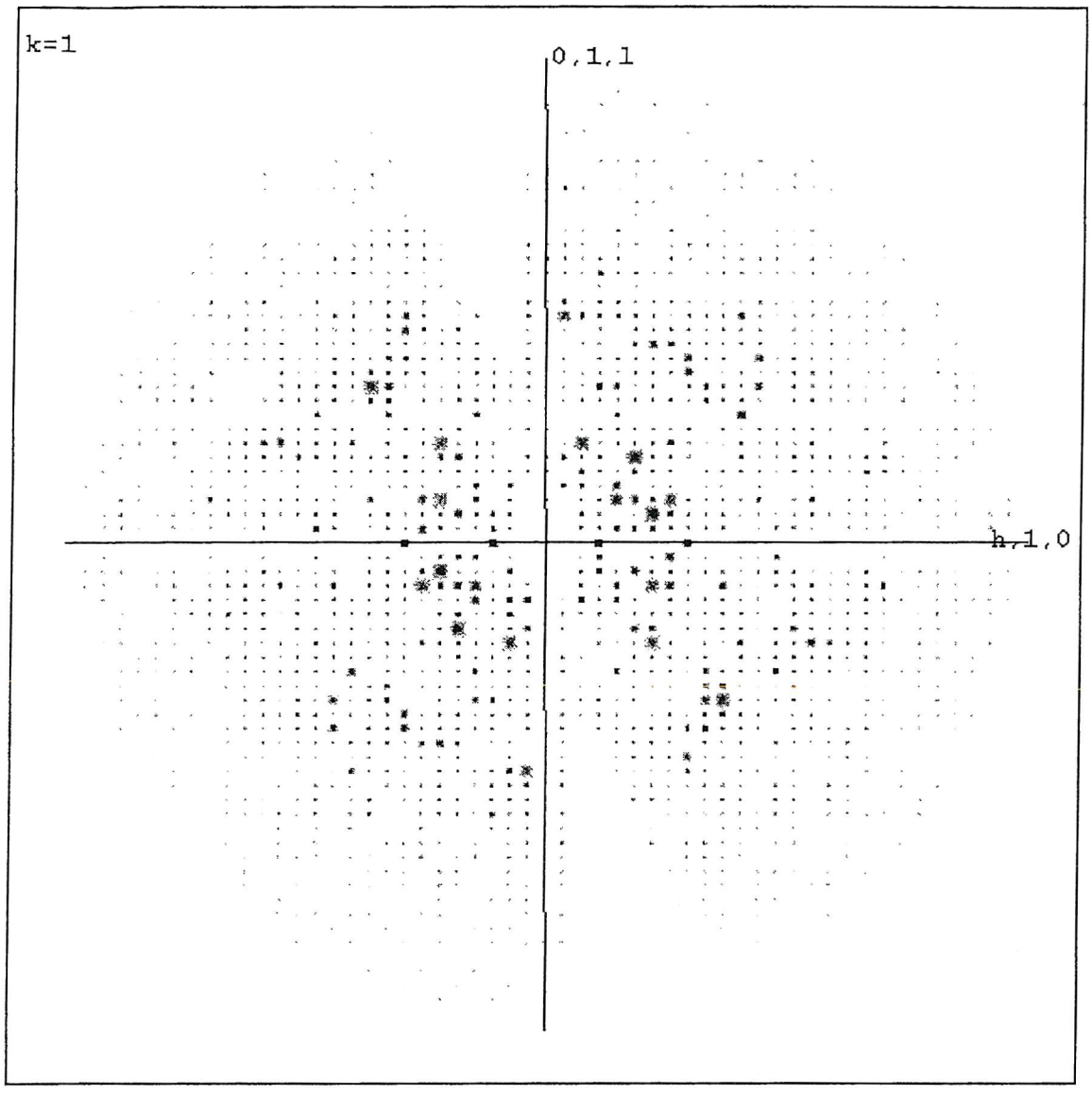


Figure 38. Pseudo-precession picture output from HKLVIEW showing the selenomethionine MhpC data set spots exhibit a symmetry pattern corresponding to primitive monoclinic rather than primitive orthorhombic. Whereas P222 data has symmetry-related spots in positions:

$|F(h\ k\ l)| = |F(\bar{h}\ \bar{k}\ l)| = |F(\bar{h}\ k\ l)| = |F(h\ \bar{k}\ l)| = |F(h\ k\ \bar{l})| = |F(h\ \bar{k}\ \bar{l})| = |F(\bar{h}\ k\ \bar{l})| = |F(\bar{h}\ \bar{k}\ l)|$.

P2 data (with b as the unique axis), has symmetry-related spots in positions:

$|F(h\ k\ l)| = |F(\bar{h}\ \bar{k}\ \bar{l})| = |F(h\ \bar{k}\ l)| = |F(\bar{h}\ k\ \bar{l})| \neq |F(\bar{h}\ k\ l)|$, where $|F(\bar{h}\ k\ l)| = |F(h\ k\ \bar{l})|$.

From these rules, we can confirm that the crystals are primitive monoclinic.

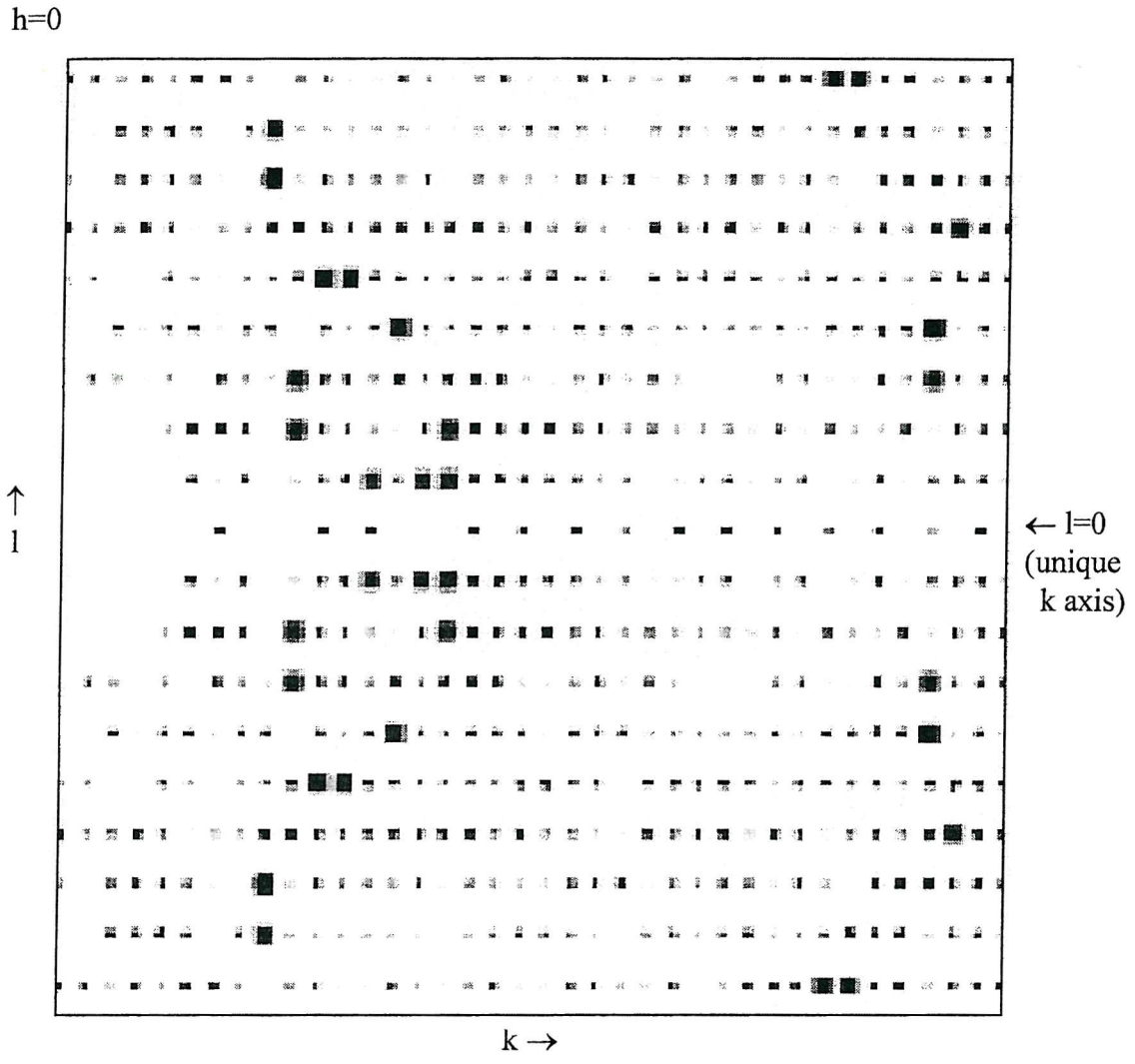


Figure 39. Pseudo-precession picture output from HKLVIEW, with data processed in spacegroup P2, showing the systematic absences along the reciprocal space lattice unique k axis, indicative of a 2_1 screw axis. Collected spots along this axis only appear at 2n positions, where n is an integer. Thus the selenomethionine MhpC crystals give rise to a 2_1 screw axis in the diffraction data.

Data processing with MOSFLM, SORTMTZ, SCALA, TRUNCATE and HKLVIEW was thus used to identify the correct space group. It was subsequently discovered that the scaled data output by this processing route was incompatible with the input required for the direct methods program SnB, which was to be used to search for the anomalous scatterers. SnB was needed to cope with the large number of seleniums present in the asymmetric unit of the protein crystal. Thus, processing had to thence be repeated using DENZO and SCALEPACK in order to overcome this problem.

4.3.9 Additional data processing of selenomethionine MhpC data sets: the DENZO route

Auto-indexing the selenomethionine MhpC M.A.D. phasing data with DENZO gave rise to the distortion index table shown in table 12. Once again, initial analysis of this table suggests that the Bravais lattice for the crystal is primitive orthorhombic. However, it is known from the aforementioned data processing described in section 4.3.8 that the Bravais lattice is actually primitive monoclinic. Moreover, the cell dimensions for the primitive monoclinic lattice favoured in the DENZO distortion index table differ from those found from MOSFLM, principally due to the b- and c-axes being swapped over (compare tables 8 and 12). Thus, DENZO was subsequently forced to use the cell parameters and spacegroup ($P2_1$) output by MOSFLM/SCALA for data processing.

Lattice	Metric tensor distortion index	Best cell (symmetrized) Best cell (without symmetry restrains)					
primitive cubic	58.21%	100.02	78.79	153.28	90.55	89.46	37.58
		115.04	115.04	115.04	90.00	90.00	90.00
I centred cubic	30.84%	168.53	153.28	163.39	21.91	96.58	88.91
		161.86	161.86	161.86	90.00	90.00	90.00
F centred cubic	35.74%	182.24	183.81	180.55	39.11	51.49	66.25
		182.20	182.20	182.20	90.00	90.00	90.00
primitive rhombohedral	9.21%	163.39	153.28	171.67	27.32	34.62	21.91
		162.95	162.95	162.95	28.97	28.97	28.97
		82.84	82.84	468.03	90.00	90.00	120.00
primitive hexagonal	10.47%	78.79	61.00	153.28	91.59	90.55	90.44
		70.46	70.46	153.28	90.00	90.00	120.00
primitive tetragonal	7.49%	61.00	78.79	153.28	90.55	88.41	89.56
		70.46	70.46	153.28	90.00	90.00	90.00
I centred tetragonal	6.39%	61.00	78.79	319.89	76.37	99.33	89.56
		70.46	70.46	319.89	90.00	90.00	90.00
primitive orthorhombic	0.86%	61.00	78.79	153.28	90.55	88.41	89.56
		61.00	78.79	153.28	90.00	90.00	90.00
C centred orthorhombic	5.81%	99.27	100.02	153.28	89.46	88.59	104.50
		99.27	100.02	153.28	90.00	90.00	90.00
I centred orthorhombic	5.81%	61.00	99.27	315.79	80.04	88.35	127.47
		61.00	99.27	315.79	90.00	90.00	90.00
F centred orthorhombic	4.30%	61.00	168.53	310.90	87.08	99.72	110.78
		61.00	168.53	310.90	90.00	90.00	90.00
primitive monoclinic	0.36%	61.00	78.79	153.28	89.45	91.59	89.56
		61.00	78.79	153.28	90.00	91.59	90.00
C centred monoclinic	5.74%	99.27	100.02	153.28	90.54	91.41	104.50
		99.27	100.02	153.28	90.00	91.41	90.00
primitive triclinic	0.00%	61.00	78.79	153.28	90.55	91.59	90.44

Table 12. Distortion index table output by DENZO during auto-indexing of the selenomethionine MhpC f ' min data set. Notice how the primitive orthorhombic Bravais lattice appears to be the optimal choice representing the crystal. In reality, the lattice type is primitive monoclinic, with the β-angle being very close to 90.0 degrees.

Once again, with DENZO, the statistical accuracy of the agreement between the diffraction spots and their predicted positions, as measured by the chi^2 values, was very good. The 'poorest' chi^2 values for each data set processed were as follows: f " max (e1): x=0.90, y=0.96; f ' min (e2): x=0.93, y=0.98 and remote (e3): x=0.87, y=0.93, where ideally these should equal 2.00 or less.

Scaling of each of the three data sets using SCALEPACK gave rise to the data presented in table 13.

Data set	Wavelength (Å)	Resolution (Å)	R_{merge} (%)	$I>3\sigma(I)$ (%)	Mean Multiplicity	Completeness (%)
Se e1, f" max	0.978689	2.30	2.9	75.6	1.3	74.7
Se e2, f' min	0.979035	2.30	3.1	73.7	1.3	74.6
Se e3, Remote	0.8855	2.30	3.0	77.7	1.3	81.3

Table 13. Statistics derived from SCALEPACK output displaying key features of the scaled selenomethionine MhpC X-ray data at each wavelength collected. It is important to be aware that in scaling each data set, the Friedel pairs were kept separate from one another, i.e. they were not merged (due to the anomalous differences between them).

The statistics presented in table 13 derive from scaled selenomethionine MhpC data collected at each of the three wavelengths selected for M.A.D. phasing. During scaling, Friedel pairs were kept as distinct diffraction spots and thus were not merged together. This accounts for features of the data such as the apparently low completeness, R_{merge} and multiplicity values.

A second round of SCALEPACK at each wavelength using the output data obtained from the first round just described, was performed to detect the presence and strength of the anomalous signal in the data. In this round, no scaling is performed, but merging of Friedel pairs does now take place. The results from this are summarised in tables 14 and 15.

Data set	Wavelength (Å)	Resolution (Å)	R_{merge} (%)	$I>3\sigma(I)$ (%)	Mean Multiplicity	Completeness (%)
Se e1, f" max	0.978689	2.30	5.5	76.4	1.6	90.2
Se e2, f' min	0.979035	2.30	4.0	74.7	1.6	90.2
Se e3, Remote	0.8855	2.30	4.0	80.2	1.7	93.1

Table 14. Summary of the statistics output from the second round of SCALEPACK runs. Both Friedel- and Bijvoet pairs were involved in merging of the data in this instance.

(a) f" max data set

Shell limit	Lower Angstrom	Upper	Average I	Average error	Average stat.	Norm. Chi**2	Linear R-fac	Square R-fac
99.00	4.96		13939.1	332.0	332.0	22.712	0.057	0.061
4.96	3.93		14153.2	339.4	339.4	8.402	0.041	0.043
3.93	3.44		10402.9	286.2	286.2	5.706	0.046	0.105
3.44	3.12		5903.5	239.2	239.2	3.333	0.051	0.051
3.12	2.90		2953.5	196.3	196.3	2.148	0.068	0.065
2.90	2.73		1770.0	199.0	199.0	1.398	0.092	0.099
2.73	2.59		1244.8	195.2	195.2	1.275	0.129	0.516
2.59	2.48		990.5	195.1	195.1	1.063	0.132	0.138
2.48	2.38		854.8	239.1	239.1	0.951	0.168	0.175
2.38	2.30		708.5	287.2	287.2	0.947	0.285	0.657
All reflections			5534.6	250.6	250.6	5.686	0.055	0.072

(b) f' min data set

Shell limit	Lower Angstrom	Upper	Average I	Average error	Average stat.	Norm. Chi**2	Linear R-fac	Square R-fac
99.00	4.96		10673.8	261.8	261.8	7.180	0.033	0.036
4.96	3.93		10517.5	258.9	258.9	2.997	0.026	0.027
3.93	3.44		7588.0	221.2	221.2	2.385	0.034	0.121
3.44	3.12		4278.4	184.0	184.0	1.589	0.038	0.039
3.12	2.90		2103.3	150.2	150.2	1.259	0.057	0.054
2.90	2.73		1244.2	150.4	150.4	1.019	0.084	0.092
2.73	2.59		879.1	149.0	149.0	1.066	0.129	0.486
2.59	2.48		688.5	148.9	148.9	0.831	0.132	0.139
2.48	2.38		591.0	184.7	184.7	0.981	0.192	0.241
2.38	2.30		488.9	220.9	220.9	1.064	0.285	0.392
All reflections			4093.1	192.8	192.8	2.279	0.040	0.066

(c) remote data set

Shell limit	Lower Angstrom	Upper	Average I	Average error	Average stat.	Norm. Chi**2	Linear R-fac	Square R-fac
99.00	4.96		18507.0	341.1	341.1	14.990	0.038	0.042
4.96	3.93		19417.0	355.5	355.5	5.514	0.027	0.030
3.93	3.44		13916.0	305.4	305.4	4.123	0.031	0.039
3.44	3.12		7890.7	266.7	266.7	2.517	0.040	0.144
3.12	2.90		3834.6	228.5	228.5	1.622	0.052	0.049
2.90	2.73		2302.7	219.5	219.5	1.434	0.075	0.072
2.73	2.59		1662.8	221.1	221.1	1.268	0.116	0.834
2.59	2.48		1331.5	220.5	220.5	1.027	0.107	0.115
2.48	2.38		1129.3	228.4	228.4	0.937	0.122	0.112
2.38	2.30		886.6	235.6	235.6	0.961	0.165	0.191
All reflections			7333.3	263.7	263.7	3.795	0.040	0.067

Table 15. Extracts from the second round SCALEPACK output files showing, of particular note, the Chi^2 values representing the anomalous signal in each data set. Chi^2 values greater than 1.0 may indicate the presence of an anomalous signal. Notice that the largest anomalous signal comes from the f" max data set (a) which was collected at the peak of the selenium absorption edge for MhpC, where anomalous scattering is at its greatest.

The output from the second SCALEPACK round is not used as direct input to any further program, but the chi^2 values for the f " max data set shown in table 15, provide a useful guide to determine what data resolution should be used for finding the anomalous scatterer sites in SnB. In other words, the resolution shells where significant and useful anomalous data is present are revealed for subsequent analysis. From table 15 (a), the usable anomalous data, represented by a chi^2 value >1.0, stretches to about 2.6Å at best. Notice that with table 14, when compared with table 13, how the data becomes more complete when Friedel pairs are merged together. This is due to absent spots in the data being accounted for by their Friedel partner. This is relevant to this work because it shows that more complete data has been collected than appears to be the case from table 13, which thus ensures a detailed and accurate refinement of the final structure can be achieved.

The mean multiplicity value increases from table 13 to table 14, since in the latter, the Friedel partners are regarded as additional copies of the same spots for merging.

Finally, the R_{merge} values increase from table 13 to table 14 since the Friedel pairs are merged in table 14 once again, which thus means that averaged intensities for a given Friedel pair, where these are merged, will be divergent from their initial intensity values due to their anomalous differences, and thus the greater the R_{merge} values will be in this instance since the R_{merge} is a measure of how close in intensity symmetry-related spots are.

4.3.10 Determining the selenomethionine positions

The scaled f" max data set with the Friedel pairs unmerged was next input into DREAR. The list in table 16 shows a sample of the normalised anomalous difference values (labelled "diffE") output by the program. It was from these diffE values that the selenomethionine positions within the unit cell were subsequently calculated.

h	k	l	E1	sig(E1)	E2	sig(E2)	diffE	sig(diffE)	rank
1	32	12	0.413	0.054	0.783	0.112	4.162	1.399	1
3	4	4	0.480	0.010	0.717	0.015	3.753	0.286	2
0	10	11	0.520	0.011	0.269	0.008	3.718	0.202	3
5	23	19	1.954	0.055	1.557	0.069	3.705	0.824	4
-11	24	13	1.188	0.042	1.556	0.082	3.557	0.890	5
-3	8	1	0.633	0.015	0.851	0.017	3.456	0.360	6
-4	41	2	0.913	0.077	1.245	0.164	3.449	1.882	7
1	8	1	0.941	0.018	0.729	0.013	3.380	0.355	8
11	34	6	0.432	0.051	0.788	0.069	3.361	0.810	9
4	38	8	0.352	0.063	0.038	0.006	3.275	0.660	10
-3	7	3	0.762	0.017	0.558	0.011	3.228	0.321	11
17	18	7	0.768	0.082	0.393	0.079	3.178	0.965	12
-4	18	12	0.355	0.009	0.588	0.014	3.149	0.225	13
-2	35	12	0.852	0.106	1.151	0.225	3.097	2.576	14
10	25	8	1.451	0.044	1.188	0.034	3.067	0.648	15
2	11	3	0.524	0.010	0.334	0.007	2.991	0.193	16
-3	31	7	0.455	0.011	0.687	0.013	2.945	0.216	17
-4	37	13	0.428	0.041	0.752	0.077	2.944	0.793	18
15	4	15	0.850	0.108	0.514	0.084	2.893	1.178	19
12	2	19	0.578	0.069	0.913	0.078	2.868	0.892	20
4	20	18	0.620	0.048	0.351	0.048	2.862	0.722	21
3	30	17	0.684	0.045	0.988	0.089	2.836	0.930	22
-2	16	2	1.004	0.014	1.186	0.017	2.819	0.341	23
11	16	8	1.418	0.046	1.194	0.038	2.819	0.751	24
8	1	21	1.565	0.140	1.263	0.079	2.805	1.493	25
-11	17	12	1.023	0.029	1.271	0.036	2.780	0.518	26
-4	13	5	0.141	0.009	0.322	0.011	2.775	0.218	27
-3	8	7	1.467	0.028	1.288	0.023	2.768	0.560	28
6	1	20	0.517	0.064	0.257	0.062	2.756	0.944	29
11	14	13	1.250	0.068	1.008	0.039	2.730	0.884	30
2	3	3	0.507	0.010	0.336	0.008	2.728	0.205	31
-10	6	5	0.581	0.017	0.390	0.011	2.725	0.289	32
1	26	12	1.004	0.016	0.788	0.014	2.720	0.268	33
4	5	4	0.408	0.013	0.237	0.013	2.691	0.290	34
8	14	3	1.039	0.030	0.856	0.022	2.681	0.545	35
-5	11	3	1.240	0.024	1.413	0.026	2.675	0.547	36
-4	18	6	1.070	0.023	0.891	0.017	2.652	0.424	37
10	5	21	0.674	0.114	0.356	0.100	2.648	1.263	38
2	28	19	1.421	0.058	1.121	0.080	2.642	0.870	39
2	14	1	0.728	0.013	0.560	0.009	2.629	0.248	40

Table 16. Extract from a long list output from DREAR, showing the strongest anomalous difference values present in the scaled f" max data set. The Miller indices h, k and l define the integers representing a particular spot and its Friedel mate, whose intensity difference is to be measured. "E1" and "E2" are the normalised anomalous intensity values for each member of a Friedel pair. "SigE1" and "sigE2" indicate the degree of error associated with each of the "E1" and "E2" measurements, respectively. "DiffE" and "sigdiffE" are the normalised anomalous difference value for a given Friedel pair, and the associated error, respectively.

SnB performed the calculations to determine the anomalous scatterer positions, and the results are displayed in figure 40, table 17 and in figure 41.

Figure 40 displays a histogram indicating how successful each of the SnB trials is likely to be, as expressed by the R_{min} values. The lower the R_{min} value, the greater the probability of a correct solution being found for the anomalous scatterer positions.

Table 17 shows the best list of possible selenomethionine positions from the numerous trials run, which are expressed as fractional coordinates in three dimensional space within the unit cell.

Figure 41 shows how the R_{min} value varied with each SnB cycle for the most successful trial. This trial corresponds to that which yielded the possible anomalous scatterer positions shown in table 17.

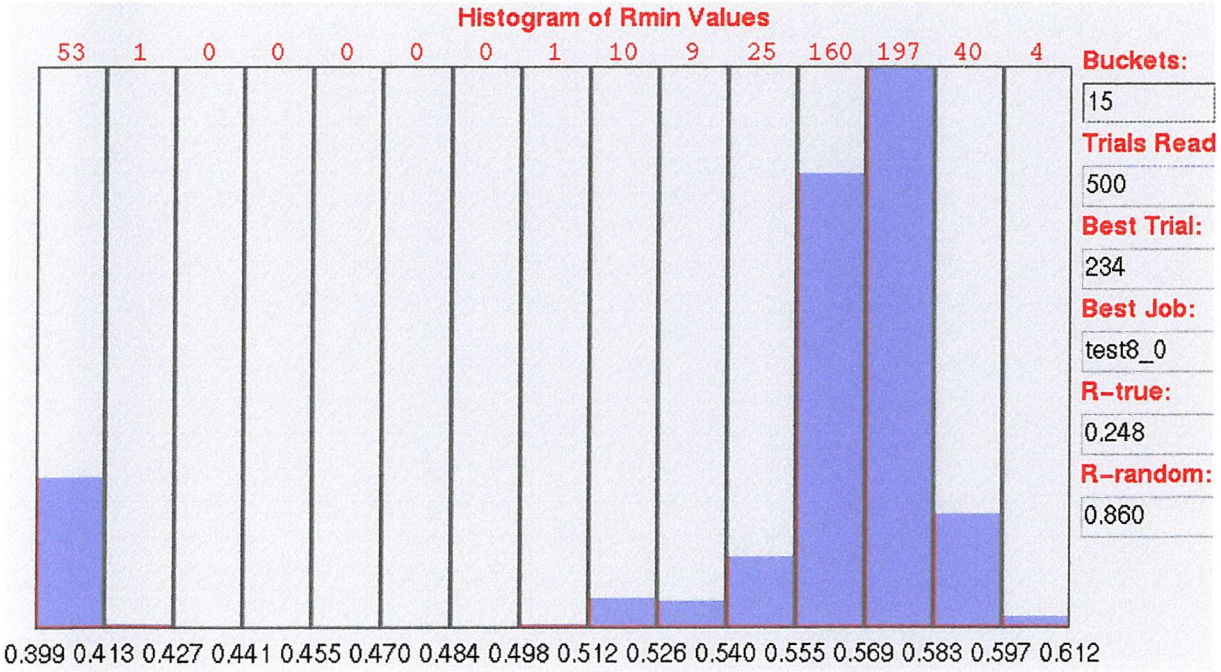


Figure 40. Histogram output by SnB indicating the relative success of each of the SnB trials as expressed by the R_{min} values. Note that 53 of the 500 trials gave solutions with corresponding Rmin values of between 0.399 and 0.413, which is in distinct contrast with the majority of trials, which have R_{min} values of around 0.5-0.6.

x	y	z	Peak size	Rank
0.683527	0.913006	0.684716	19.11	1
0.323117	0.465407	0.815227	18.86	2
0.564489	0.985293	0.517882	18.22	3
0.538764	0.968643	0.550040	17.37	4
0.093569	0.040873	0.567720	15.73	5
0.878275	0.316140	0.694261	14.95	6
0.456824	0.405374	0.682492	14.43	7
0.309763	0.075824	0.639056	14.42	8
0.481349	0.430366	0.881584	14.17	9
0.423565	0.389292	0.645243	13.77	10
0.216070	0.929832	0.569079	13.71	11
0.742335	0.296545	0.565242	13.71	12
0.391933	0.900204	0.526058	13.26	13
0.613265	0.465104	0.653260	13.16	14
0.548929	0.850230	0.638471	13.13	15
0.435332	0.427902	0.968339	13.03	16
0.247068	0.954396	0.559165	12.77	17
0.471812	0.524411	0.764494	12.70	18
0.747078	0.408693	0.691063	12.70	19
0.783318	0.431446	0.690378	12.49	20
0.198058	0.080850	0.717107	12.35	21
0.766612	0.566102	0.555228	11.88	22
0.708393	0.314122	0.786684	11.78	23
0.766120	0.281030	0.839182	11.39	24
0.514232	0.942552	0.754253	11.26	25
0.561482	0.945949	0.837784	11.15	26
0.605441	0.492023	0.546273	10.35	27
0.656238	0.290917	0.767950	9.95	28
0.255539	0.049935	0.656973	9.83	29
0.467221	0.915925	0.789713	8.60	30
0.594360	0.380113	0.575895	8.57	31
0.807716	0.630328	0.586495	8.48	32
0.531387	0.455003	0.916334	8.21	33
0.375536	0.146977	0.619206	7.93	34
0.047353	0.530098	0.637316	6.39	35
0.991090	0.401854	0.961996	6.34	36
0.394803	0.330622	0.784381	6.31	37
0.223465	0.740853	0.536735	6.14	38
0.667611	0.886322	0.948935	5.96	39
0.117958	0.443072	0.610783	5.92	40

Table 17. List of possible selenium positions within the asymmetric unit of the MhpC crystal. The putative sites are expressed as fractional coordinates and are ranked according to their probability of correctness.

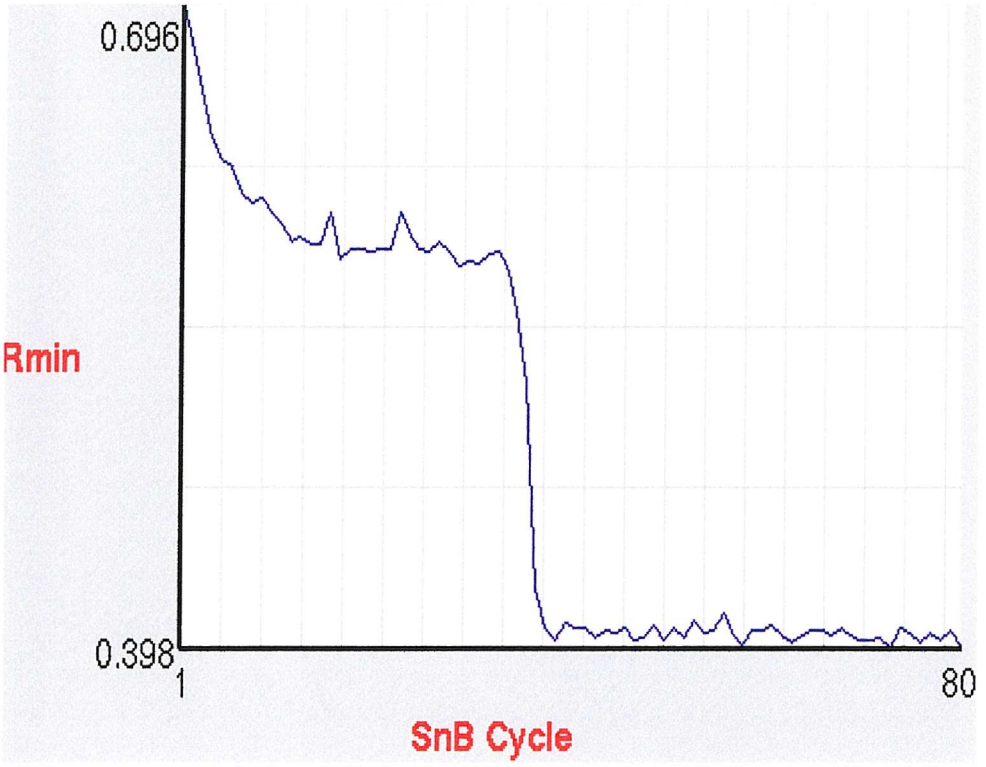


Figure 41. Plot output from SnB showing how the R_{min} value, representing the relative success of SnB, dropped during the most successful trial. This trial gave rise to the list of putative selenium sites shown in table 17.

FINDNCS was next used to pick out the non-crystallographic symmetry-related anomalous scatterer positions from the entire list of possible positions found by SnB. The data output by FINDNCS are shown in table 18. From this data, it can be seen that three distinct non-crystallographic symmetry axes have been identified. Each of these axes relates a given cluster of anomalous scatterer sites (found in one protein molecule) to each of the other clusters of non-crystallographic symmetry-related sites output by SnB. Thus, high- and low-probability anomalous scatterer sites from the list output by SnB can be distinguished on the basis of their symmetry with other sites, and hence, in all likelihood, the real anomalous scatterer positions within the crystal can be accurately chosen.

```
 NCS1     with matching pairs          32
  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16
  2  1 10  7  6  5  4 28 25  3 20 22 14 13 18 26

 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 33
 19 15 17 11 24 12 29 21  9 16 31  8 23 33 27 30

 NCS matrix:
 0.99569     -0.09203     -0.01145
-0.09203     -0.99576      0.00053
-0.01145      0.00053     -0.99993
 6.88549    138.12453     89.96854
RMS: 0.376      Screw:     0.00      Radii:    22.06
Polar angle:    89.67  177.36  180.00      &     90.33   -2.64 -180.00
Center:    38.62    67.44    44.78

-------------------------------------------------------------------------
 NCS2     with matching pairs          32
  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16
  5  6 11 17  1  2 19 30 23 20  3 18 27 31 22 24

 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 33
  4 12  7 10 26 15  9 16 29 21 13 33 25  8 14 28

 NCS matrix:
-0.99963      0.01114     -0.02486
 0.01113     -0.66589     -0.74597
-0.02486     -0.74597      0.66552
77.59825    145.32336     66.24689
RMS: 0.545      Screw:     0.00      Radii:    22.06
Polar angle:   155.86   88.09  180.00      &     24.14  -91.91 -180.00
Center:    38.62    67.44    44.78

-------------------------------------------------------------------------
 NCS3     with matching pairs          32
  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16
  6  5 20 19  2  1 17 33 29 11 10 15 31 27 12 21

 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 33
  7 22  4  3 16 18 25 26 23 24 14 30  9 28 13  8

 NCS matrix:
-0.99606      0.08086      0.03628
 0.08086      0.66160      0.74548
 0.03628      0.74548     -0.66554
70.01953    -13.68632     22.91033
RMS: 0.557      Screw:     0.00      Radii:    22.06
Polar angle:   114.14  -92.79  180.00      &     65.86   87.21 -180.00
Center:    38.62    67.44    44.78
```

Table 18. Key output from FINDNCS, revealing three non-crystallographic symmetry axes, which define how each of the clusters of selenium sites, corresponding to individual MhpC protomers, are symmetrically-related to one another within the asymmetric unit. Each non-crystallographic symmetry axis, labelled "NCS1", "NCS2" OR "NCS3", is presented as both a non-crystallographic symmetry matrix, and as a series of angles defining the axis about a specified centre point in the unit cell. The pairs of numbers labelled 1 to 33 represent pairs of selenium sites which are symmetry-related due to either NCS1, NCS2 or NCS3. The particular selenium sites that the numbers 1 to 33 correspond to are shown in table 17. Note that the putative selenium site labelled number 32, in table 17 does not share symmetry with any of the other putative sites with the symmetry operators NCS1, NCS2 and NCS3, whereas site number 33 does in each case. How closely the clusters of selenium sites are symmetry-related to one another is indicated by the root-mean-squared (RMS) difference between them when they are best overlaid on top of each other after rotation.

4.3.11 Calculation and refinement of protein phases and modelling of the MhpC structure

Once all of the selenium sites had been located, the CNS program MAD_PHASE next used this information to calculate experimental phases for the complete diffraction data. The accuracy of the experimental phase information is indicated by statistics output by MAD_PHASE. These are shown in table 19 below. The "phasing power" represents how strong the overall signal attributed to the anomalous scatterers is compared with the associated statistical error in the measurement of the anomalous signal. This indicates how useful the anomalous signal is for determining experimental phases. The "figure of merit" is also important, since it is an indicator of how likely the experimental phases are to be correct, i.e. the figure of merit gives an idea of the error associated with the phase estimates for each of the individual diffraction spots.

```
                                    phasing power
                                    -------------
phasing statistics of F_E3_FRIED (LOC target working set, excluding outliers)
 #bin | resolution range | #refl |   all
    1   4.57   24.92       7772      2.3887
    2   3.63    4.57       6739      1.6630
    3   3.17    3.63       6026      1.1995
    4   2.88    3.17       4548      0.8994
    5   2.68    2.88       3348      0.6319
    6   2.52    2.68       2408      0.4819
    7   2.39    2.52       1438      0.4215
    8   2.29    2.39         62      0.2995
 #bin | resolution range | #refl |
    1   2.29   24.92      32341      1.3922
.....................................................................
                                    phasing power
                                    -------------
phasing statistics of F_E3_FRIED (LOC phase set, excluding outliers)
 #bin | resolution range | #refl |   all
    1   4.57   24.92      13316      2.3675
    2   3.63    4.57      13350      1.6191
    3   3.17    3.63      12680      1.1523
    4   2.88    3.17      11624      0.8789
    5   2.68    2.88      10498      0.6090
    6   2.52    2.68       9252      0.4711
    7   2.39    2.52       8265      0.4086
    8   2.29    2.39       6558      0.2977
 #bin | resolution range | #refl |
    1   2.29   24.92      85543      1.0851
.....................................................................
                                    figure-of-merit
                                    ---------------
Figure of merit for global phase set
 #bin | resolution range | #refl |   all
    1   4.57   24.92      14657      0.8414
    2   3.63    4.57      14770      0.7529
    3   3.17    3.63      14386      0.6541
    4   2.88    3.17      13680      0.5370
    5   2.68    2.88      12890      0.4131
    6   2.52    2.68      11982      0.3203
    7   2.39    2.52      11055      0.2383
    8   2.29    2.39       9174      0.1197
 #bin | resolution range | #refl |
    1   2.29   24.92     102594      0.5176
```

Table 19. Statistics output by the CNS program MAD_PHASE, used to indicate how accurate the experimental protein phases are which are generated at this stage. The phasing power tables show how strong the anomalous scattering signal in the overall data is in various resolution shells. Phasing power values are presented in the columns labelled "all", which represent information from all of the collected diffraction spots in certain resolution shells. Where phasing power values exceed 1.0, the anomalous signal is strong, and thus particularly useful information can be derived from diffraction spots in the corresponding resolution shells. The figure of merit table shows how accurate the experimental phases are in several resolution shells. The values are expressed as fractions, where a figure of merit of 1.0 corresponds to a set of ideal phase values.

Experimental protein phases are calculated separately from the two enantiomeric arrangements of the selenium site positions. In order to resolve the selenium sites' hand ambiguity, two electron density maps are generated using the experimental protein phases output using each set of sites. Where the selenium site positions adopt the correct enantiomeric arrangement, the resultant electron density map is generated from non-random phase information, and as such, shows evidence of a partition within the density between protein and solvent. In contrast, where the selenium sites present in the incorrect enantiomeric configuration are used to generate protein phases, the resultant map has a random distribution of electron density due to the phase information being incorrect. Sections of the two electron density maps generated from each of the enantiomeric arrangements of selenium sites are shown in figure 42, for comparison.

(a) (b)

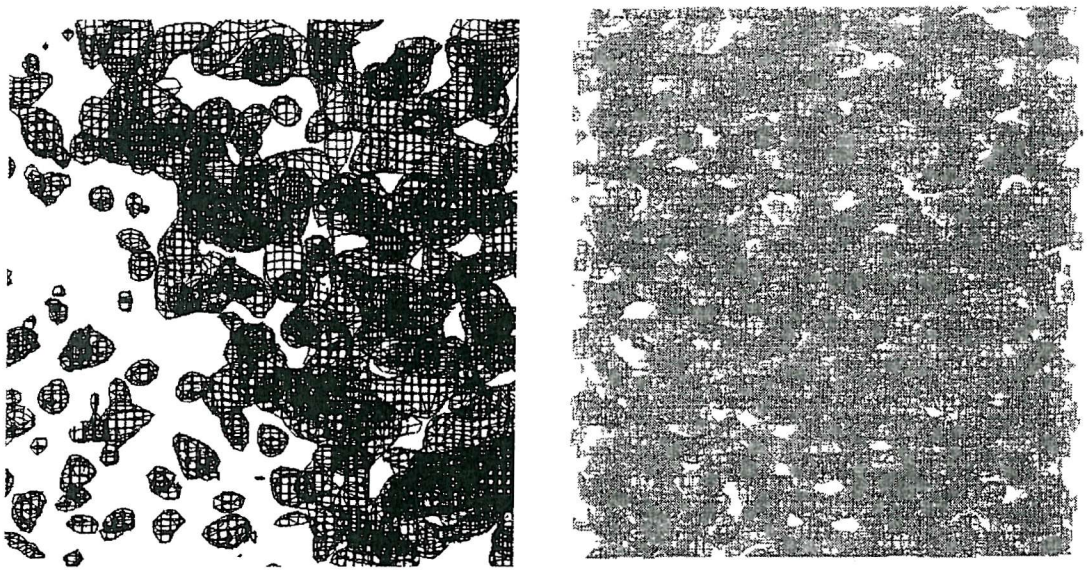


Figure 42. Electron density maps of the same region of three dimensional space within the MhpC unit cell, where phases used in map generation were derived from a different enantiomeric arrangement of selenium sites in each case. Comparison of the two maps allows elucidation of the correct set of experimental phases for the MhpC structure, and also reveals the correct enantiomeric arrangement of selenium sites. (a) Map showing evidence of a protein-solvent boundary occurring within the electron density, indicative of a correct set of experimental protein phases. Note how the bottom-left portion of the density is relatively sparse (corresponding to solvent), whilst the top-right portion is dense (corresponding to protein). (b) Map of the same region showing random density evenly distributed throughout the picture. Thus, the phases used in the creation of this map are incorrect.

Density modification using the correct set of experimental protein phases along with their associated structure factor amplitudes was next employed to generate a refined set of phases and amplitudes giving rise to electron density clearly defining the protein structure (i.e. the background noise had been effectively removed in the resultant electron density map). An estimate of the solvent content of the crystal needed to be provided in order to perform the density modification, and the CNS program MATTHEWS_COEF was used to obtain this. The solvent content was estimated at 59%.

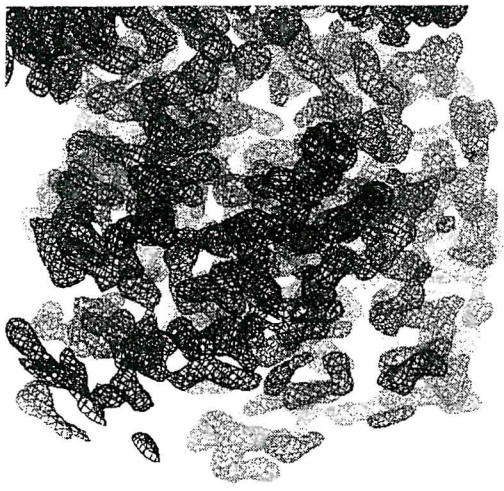
The final statistics output from the density modification process are shown in table 20. Examining the statistics output, the effect of the density modification can be clearly seen in the greatly improved figure of merit values compared with those output during the initial calculation of experimental phase values (see table 19).

#bin	Resolution range		#refl	FOM	dphi
1	8.47	100.00	2616	0.8709	2.6374
2	6.72	8.47	2600	0.9592	3.8818
3	5.87	6.72	2655	0.9578	4.0238
4	5.34	5.87	2584	0.9629	4.5726
5	4.96	5.34	2625	0.9536	3.8829
6	4.66	4.96	2649	0.9537	3.4667
7	4.43	4.66	2592	0.9402	3.5392
8	4.24	4.43	2622	0.9402	4.6346
9	4.07	4.24	2635	0.9455	5.2914
10	3.93	4.07	2593	0.9529	6.1647
11	3.81	3.93	2630	0.9397	5.7914
12	3.70	3.81	2583	0.9222	6.3114
13	3.60	3.70	2612	0.9285	6.7153
14	3.52	3.60	2699	0.9223	6.4580
15	3.44	3.52	2515	0.9150	6.8038
16	3.36	3.44	2659	0.9152	6.7019
17	3.30	3.36	2550	0.9187	6.9554
18	3.23	3.30	2656	0.8761	6.7833
19	3.18	3.23	2676	0.9179	7.0325
20	3.12	3.18	2562	0.8839	7.5764
21	3.07	3.12	2630	0.9103	7.6160
22	3.02	3.07	2602	0.8803	6.7710
23	2.98	3.02	2601	0.8861	7.2866
24	2.94	2.98	2649	0.8459	7.7187
25	2.90	2.94	2640	0.8833	7.1404
26	2.86	2.90	2603	0.8910	6.6369
27	2.82	2.86	2567	0.8991	7.1186
28	2.79	2.82	2636	0.8865	7.3150
29	2.76	2.79	2660	0.8510	6.7134
30	2.73	2.76	2564	0.8808	6.4409
31	2.70	2.73	2614	0.8418	7.1091
32	2.67	2.70	2647	0.8805	6.3907
33	2.64	2.67	2575	0.8378	6.5718
34	2.62	2.64	2676	0.8740	6.5129
35	2.59	2.62	2582	0.8392	6.2768
36	2.57	2.59	2621	0.8456	5.9906
37	2.54	2.57	2642	0.8030	6.6556
38	2.52	2.54	2612	0.8992	6.0284
39	2.50	2.52	2552	0.8830	5.8503
40	2.48	2.50	2650	0.8952	5.5308
41	2.46	2.48	2609	0.8590	5.8805
42	2.44	2.46	2678	0.8382	5.7184
43	2.42	2.44	2576	0.8664	5.4086
44	2.40	2.42	2564	0.7850	5.7493
45	2.38	2.40	2618	0.8233	5.5205
46	2.36	2.38	2671	0.7741	5.6346
47	2.35	2.36	2588	0.7550	6.2517
48	2.33	2.35	2600	0.8074	5.5195
49	2.32	2.33	2626	0.8625	5.9398
50	2.30	2.32	2566	0.8765	5.8276

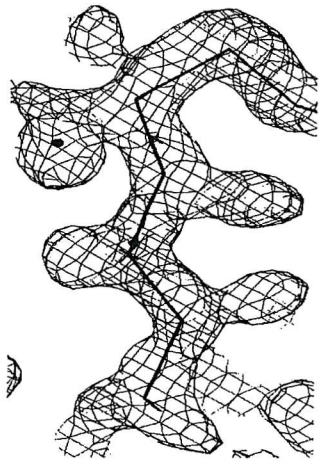
Table 20. Output from the CNS density modification program. Improved figure of merit values (labelled "FOM") are shown corresponding to diffraction data in various resolution shells.

After density modification, primary model building could thence proceed. Figure 43 below shows various images from this process.

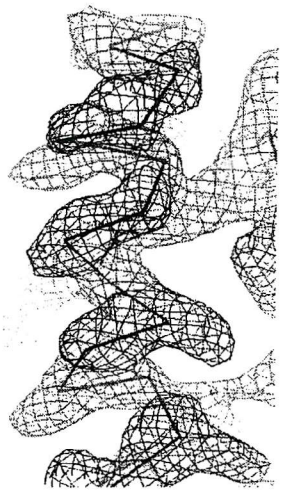
(a)



(b)



(c)



(d)

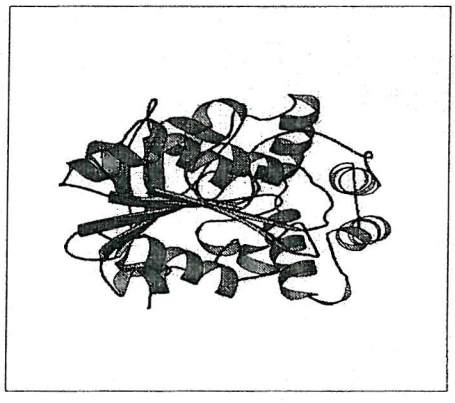


Figure 43. Various pictures tracing the development of the preliminary model of the MhpC structure. (a) Initial electron density map. (b) Beta strand with modelled Cα trace running through it. (c) Alpha helix with Cα trace modelled into it. (d) Preliminary-modelled protomer structure. Pictures (a)-(c) were generated by the graphical display package QUANTA [168], whilst picture (d) came from the molecule display package MOLSCRIPT [35].

Pictures (a)-(d) in figure 43 show how the preliminary MhpC model was constructed. Picture (a). A portion of the overall electron density is shown here before model building has begun. Several alpha helices are visible at this stage. Pictures (b) and (c). Secondary structural features showing the first stage of model building, involving 'threading' of the Cα trace through the electron density. Note the vacant side chain density projecting away from the Cα positions. Also note the dense sphere in picture (b) which corresponds to one of the selenium positions used to provide the protein phase information. Picture (d). Completed preliminary protomer structure of MhpC. It should be noted that the asymmetric unit of the protein consists of four such MhpC protomers, which are very tightly associated with one another.

The 2.3Å MhpC model structure is currently partway through refinement. The current Rfactor is 28.2, whilst the Rfree is 28.9. Water molecules have yet to be incorporated into the model. As an indicator of the success of the refinement so far, the Ramachandran plot for the current model is provided in figure 44.

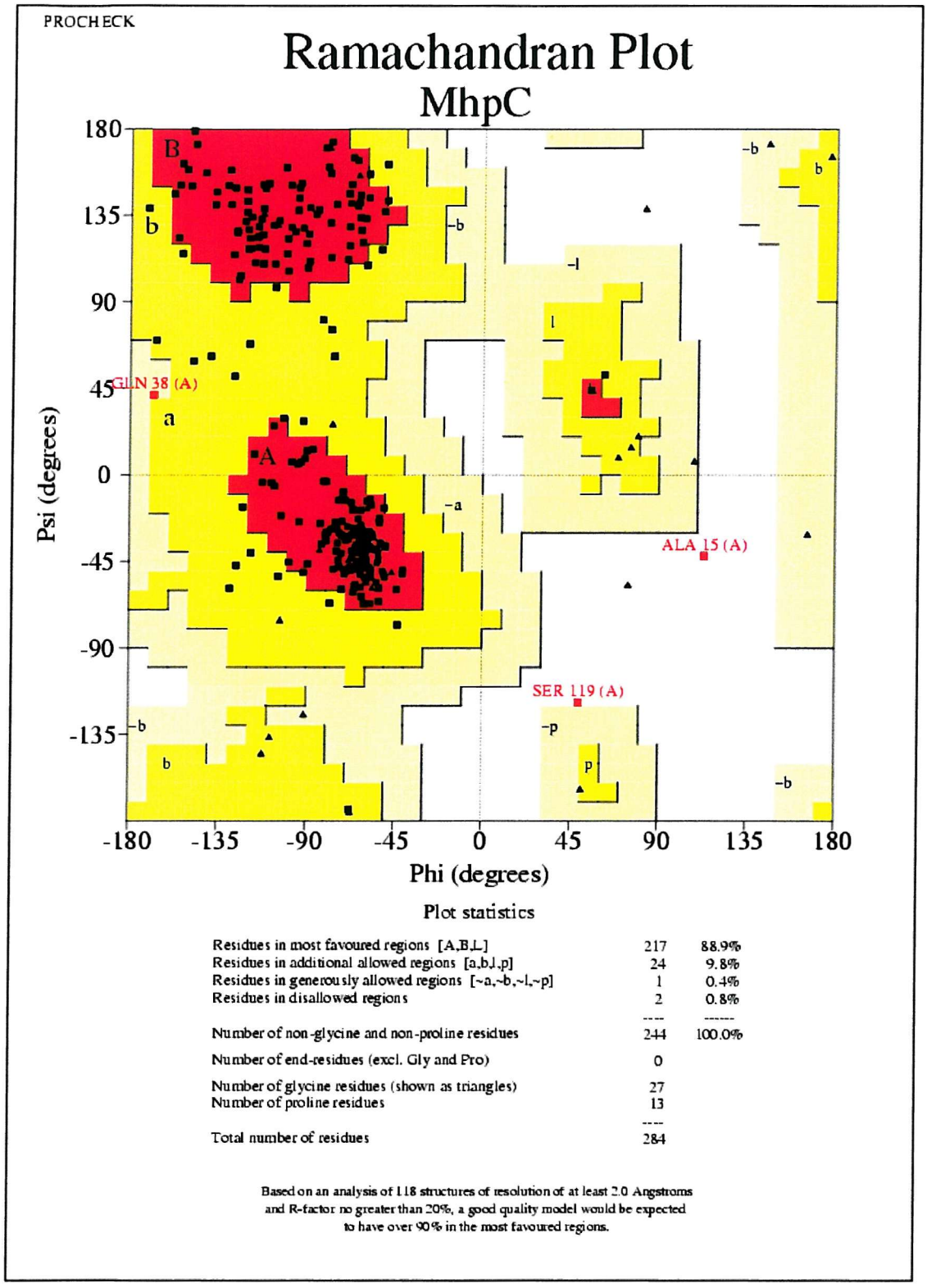


Figure 44. Ramachandran plot of the current model for an MhpC protomer. Of the residues in the disallowed regions, alanine 15 resides in a flexible N-terminal section of the model structure, and hence, appears in a diffuse region of electron density, which makes its position ambiguous in the model. Then there is serine 119, a postulated active site residue, which adopts the same sterically unfavourable conformation in other MhpC structure-related enzymes with the α/β hydrolase fold [143]. Plot output from the CCP4 structure validation program PROCHECK [167].

Chapter 5

Discussion

5. Discussion:

5.1 *Der p*1

Expression of *proDer p*1-thioredoxin fusion protein was a partial success. Protein of the correct molecular weight was produced upon induction of the *E. coli* cells, but unfortunately, the vast majority of this was channelled to an insoluble compartment within the cell, most probably to the inclusion bodies, since expression levels were reasonably high as shown by the fairly dense bands seen on the SDS-PAGE gels. The expression system itself was operating properly as shown by normal soluble thioredoxin (positive control) expression at high levels.

So why was the *proDer p*1-thioredoxin fusion protein expression being re-routed to the inclusion bodies? Firstly, *Der p*1 in the eukaryotic cells of the dust mite is a secreted protease. Thus, in the dust mite, immediately after synthesis, the protease is channelled into the endoplasmic reticulum-Golgi apparatus-vesicular system, where it undergoes disulphide rearrangements and accepts sugar attachments, amongst other things, i.e. it passes through environments distinct from that of the cytosol. So, one possible reason why the fusion protein is not expressing as a soluble entity is that within the prokaryotic *E. coli* cells in which its expression is being attempted, there is no endoplasmic reticulum or Golgi apparatus, etc., so the translated polypeptide chain here may be affected by prolonged exposure to the cytosol within the *E. coli* cell, which may lead to protein aggregation and re-routing to the inclusion bodies. Secondly, the fusion protein is approximately three times larger than the correctly folded thioredoxin control, and, as such, may be much more tricky to fold up correctly and so once again aggregation may result from this.

Perhaps fusion protein expression is achievable with this construct. Future trials should involve modification of parameters such as reduced temperature, altered oxygen levels and varied volume of growth medium, etc.

Regarding the refolding of insoluble *proDer p*1 fusion protein, the process appears to have been successful, based on the observation that soluble protein of approximately the correct molecular weight for the fusion protein had been obtained after the refolding protocol had been run. However, the efficiency of the refolding procedure seems to be quite low. Since oxidised glutathione is expensive and is used in relatively large quantities per refolding run, for relatively little return of protein, even before purification has begun, then it is unfortunately not a cost-effective way of expressing the *Der p*1, if indeed the refolding has worked correctly. Thus, again, the solution to the expression problem here must lie within the expression stage of the process rather than with the refolding stage.

Finally, autocatalysis with the *proDer p*1-thioredoxin fusion protein was unsuccessful. The simplest explanation for this is that the refolding step, whilst yielding soluble protein, may not have produced fully refolded protein which could thus not autocatalyse itself into the mature enzyme. Alternatively, of course, although *Der p*1 is homologous to its cysteine protease counterpart papain, which is capable of autocatalysis under the conditions found by Taylor, *et al* [183], it may simply be that *proDer p*1 is not capable of autocatalysing its own mature form's generation. One final hypothesis, although unlikely, is that the thioredoxin in the fusion protein may, in some way, for example through steric hindrance, interact with its *Der p*1 fusion partner, and in so doing, block autocatalysis.

5.2 Momordin

The outcome of the momordin-βNADH work was very interesting. To reiterate, the 2Å electron density ultimately produced from the crystals soaked in βNADH unveiled a region at the active site which more closely resembled nicotinamide than adenine. This came as a surprise. Before discussing this result however, it is important to briefly mention a couple of 'historical' experiments associated with this result which are relevant to this discussion. First there was the work by Xiong, *et al*, who obtained X-ray crystallographic data from crystals of the RIP trichosanthin and NADP [126]. The result here was that the NADP bound to trichosanthin remained uncleaved.

Following on from this experiment workers in Professor Steve Wood's laboratory at Birkbeck college (unpublished) unveiled a structure from crystals of momordin soaked in $NADP^+$, where the nicotinamide group was, surprisingly, cleaved in preference to the adenine. Thus, in the experiment presented in this work, we were taking this one step further: It had been assumed that the reason the nicotinamide was cleaved in preference to the adenine in that experiment may have been due to the presence of the 2' phosphate at the adenine-end of the molecule which may have prevented proper binding of the adenine-end to the enzyme via steric hindrance. Thus, the experiment within the work presented here, using βNADH as a substrate, was designed firstly, via the absent 2' phosphate, to remove any potential steric hindrance effects from occurring which should thus allow adenine access to the momordin active site, and secondly, the experiment used reduced βNADH, in order to remove the positively charged nicotinamide region from the molecule which might otherwise favour nicotinamide group cleavage through a mirroring of the cationic portion of the adenine-ribose transition state in the native reaction. Thus, the intention of the work presented using βNADH

here was to see if the adenine would now be cleaved rather than the nicotinamide. However, as we have seen, this was not the case. So why did the nicotinamide become cleaved from the βNADH substrate? It is unclear from this purely crystallographic, qualitative experiment how much adenine is released from the reaction of momordin with βNADH, so we do not know which side of the molecule is the preferred substrate, but either way, the final result is the same, i.e. the enzyme can cleave the nicotinamide from the βNADH. If the nicotinamide-end of βNADH is, however, the preferred substrate for momordin, then this may be due to the nicotinamide group being more flexible than the double ringed adenine, which might allow the nicotinamide to more easily find its way into the active site than adenine.

An alternative explanation for the nicotinamide cleavage event is that βNADH may not actually be a substrate for the enzyme at all, but that traces of contaminating cationic βNAD^+ may be present in the sample which could lead to preferential cleavage of nicotinamide from βNAD^+ via its resembling the native adenine-ribose transition state. Since once the reaction has ended, the nicotinamide remains in the active site, irreversibly trapped, then only trace amounts of βNAD^+ would be sufficient to block all of the momordin active sites, rendering the enzyme unresponsive to further substrate.

One important question is raised from this observation of momordin's ability to cleave a nicotinamide group from a nicotinamide adenine dinucleotide substrate. In addition to its known specific *N*-glycosidase activity, it is conceivable that momordin could perhaps also perform an ADP-ribosylase function. By removing the nicotinamide group from an NAD^+ substrate, for example, it could bring about ADP-ribosylation of various host cell proteins. The best known example is the eukaryotic elongation factor, eEF-2, ADP-ribosylation of which inhibits eukaryotic translation. This is an intriguing possibility that should be examined in the future. One possible way of doing this could be to add momordin to cell free extracts,

and then to purify out and analyse on a mass spectrometer, any eEF-2 present in the extract to see if any molecules had been covalently modified via ADP-ribosylase activity.

Finally, the nicotinamide itself is worthy of our attention. Why does it irreversibly block the active site? It is possible that the nicotinamide could resemble a transition state of the adenine-ribose substrate since it is a charged group of similar structure to adenine. Transition states in reactions are well documented as exhibiting stronger binding characteristics than the substrates involved, and as such, transition state analogues are potent inhibitors of enzymes due to the tightness of binding. This evidence supports the idea that nicotinamide is a transition state analogue. Nicotinamide then, through its irreversible blocking of the momordin active site may supply more information in the future regarding the design of effective inhibitors to RIPs.

This new data, along with that from the work on momordin-$NADP^+$ and from the work on trichosanthin-NADP, will ultimately have important implications regarding the mechanism of RIP catalysis. But for the moment, new compounds must be examined with the RIPs in order to answer yet more questions, until a stage is reached where sufficient data has accumulated which will allow a clear picture of the overall molecular chain of events during catalysis to be finally unveiled. Figure 45 shows a mechanism by which momordin could cleave NADH.

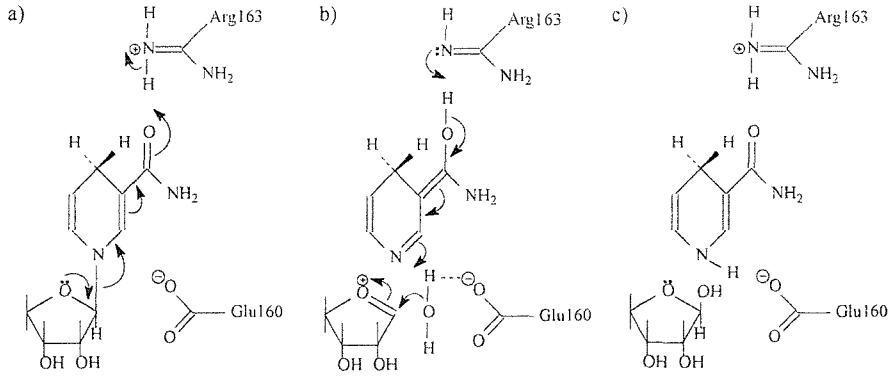


Figure 45. Putative mechanism by which momordin cleaves βNADH. a) Protonation of the nicotinamide amide group by Arg-163 leads to glycosidic bond cleavage and oxycarbonium ion formation. This is stabilised by Glu-160. b) Glu-160 assists the withdrawal of a proton from a water molecule. This allows hydroxyl attack on the ribose C1', thus fully stabilising the oxycarbonium ion intermediate. Meanwhile, the proton withdrawn from the water molecule is transferred to the nicotinamide and simultaneous electron shifts mediated by Arg-163 lead to product formation. c) The catalytic side chains return to their initial states and the products are released.

5.3 MhpC:

A descriptive journey around the MhpC structure

MhpC is a newly found member of the α/β hydrolase fold family of hydrolytic enzymes [143]. These share a similar core composed of an α/β sheet of eight β-strands linked by α-helices, and all possess a catalytic triad of active site residues which adopt an homologous arrangement and reside on highly conserved loops [143]. What follows here is a descriptive journey around the MhpC protomeric structure, which is intended to provide a fuller appreciation of the overall arrangement of the enzyme's α/β hydrolase fold. Figure 46, and the pictures shown in figure 47 (a)-(w), are provided as an additional guide to help visualise the structure.

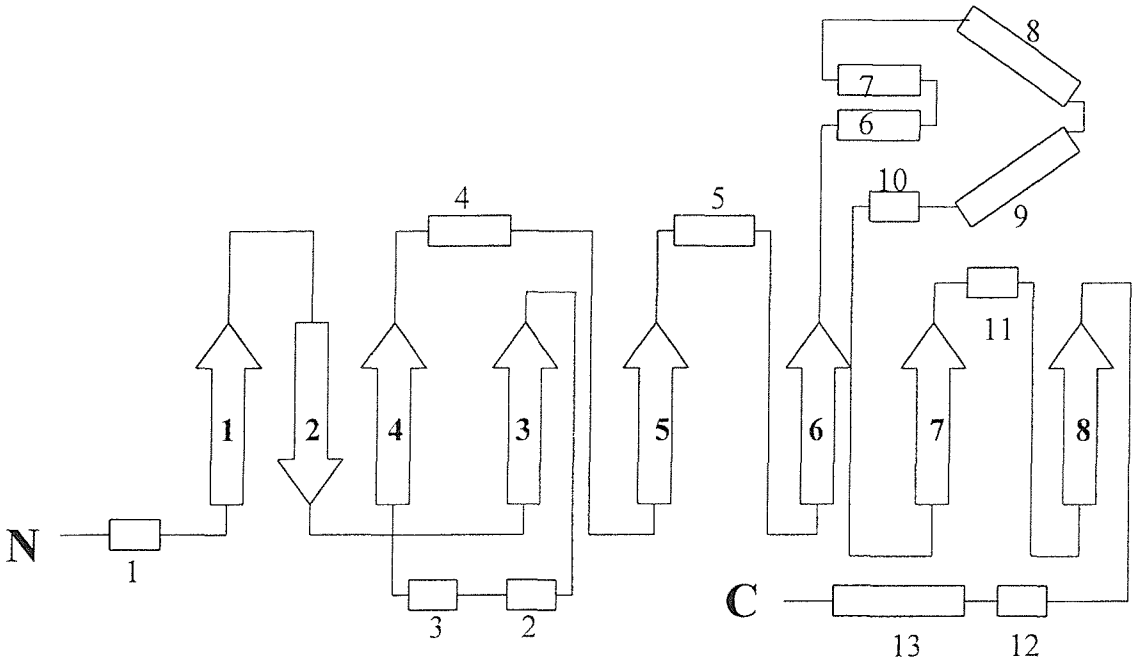


Figure 46. Diagrammatic overview of the MhpC α/β hydrolase fold. Arrows represent β-strands, whilst boxes denote α-helical regions. "N" and "C" are the two termini of the protein.

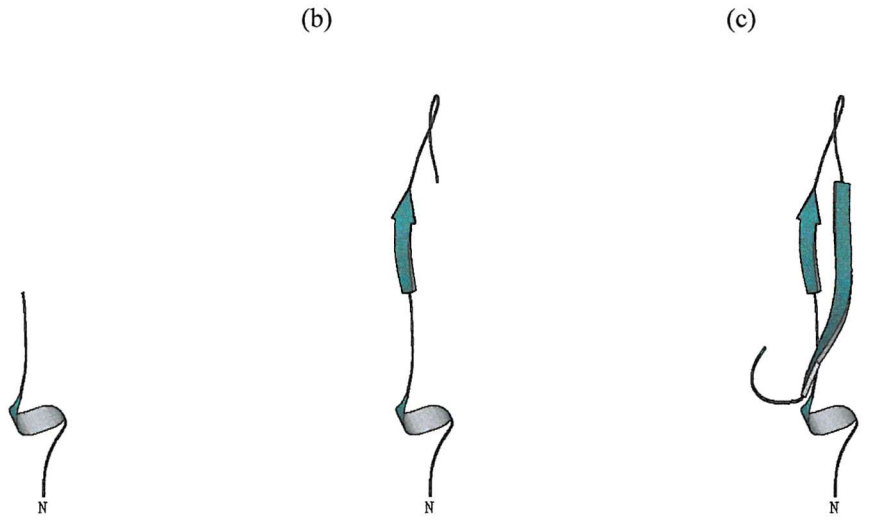


Figure 47. A diagrammatic journey around the MhpC structure. Coils represent α-helical regions, whilst arrows denote β-strands. "N" and "C" are the two termini. All pictures were generated from MOLSCRIPT [35]. (a) The N-terminus and α-helical region 1. (b) β-strand 1. (c) β-strand 2.

(a) Picture yourself at the N-terminus of the protein. This is equivalent to standing atop the "N" in the diagram. From this point, we can now journey around the structure at the molecular level, and analyse the layout of the protein in detail as we go.

The first small stretch seems fairly straight and simple, but then, almost immediately, we find ourselves passing through our first sample of secondary structure, which is just a small helical turn that gently rocks us around through about 270° or so, and then leads us into another very straight section. (b) After a very short period, we start to shift slightly from side to side in a zigzag motion and notice we are travelling along our first β-strand, which is quite a short stretch of four amino acids, and which kinks slightly, so as to rotate us round just a fraction. Upon leaving the β-strand, we next enter a reasonable-sized hairpin loop region, which sends us more or less back on ourselves, heading towards the N-terminus. (c) The zigzagging motion then returns as we enter a second β-strand, which packs very close to the first. This one is twice as big as the first one, and it rotates us around by 90°, before passing us into a semicircular loop.

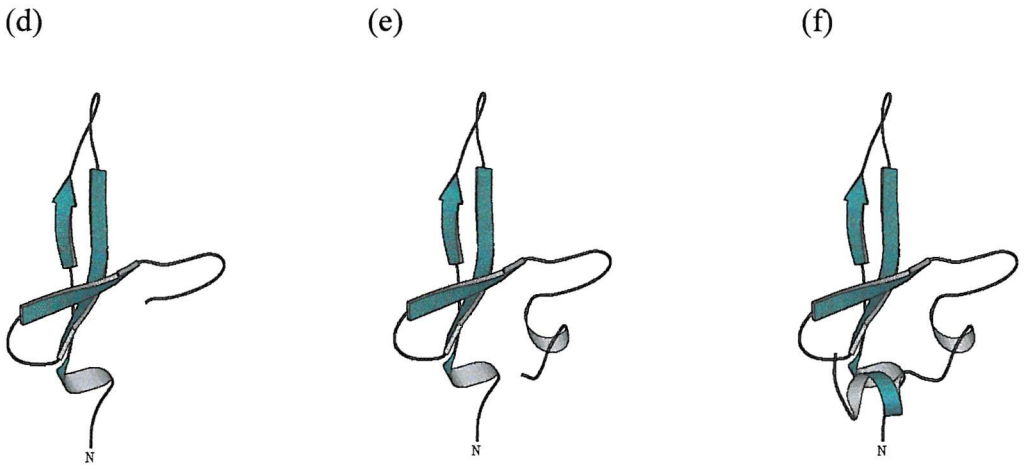


Figure 47. (d) β-strand 3. (e) α-helical region 2. (f) α-helical region 3.

(d) At the loop's end, we find ourselves passing along a third β-strand, which runs in an almost perpendicular direction to the rest of the molecule we've journeyed through so far. From here, we are next propelled through a sharp hairpin loop, and then (e) spin through a second ¾ α-helical turn and hump. The helical turn and hump adopt a very similar appearance to the region from N-terminus to the first loop, and these two regions align themselves reasonably close to each other at this stage. (f) After the hump, we pass through a short 450° coil of α-helix. The axis running through this helical region runs parallel with the third β-strand, and the short loop which emerges from the helical turn, heads off so that we are now sitting directly adjacent to the N-terminal end of the third β-strand.

(g) (h)

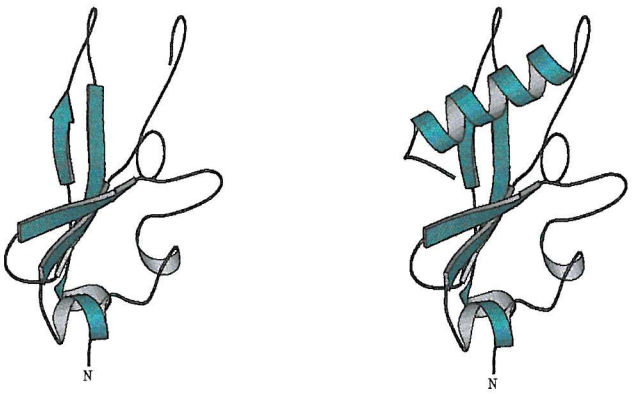


Figure 47. (g) β-strand 4. (h) α-helical region 4.

(g) We next find ourselves zigzagging along a fourth β-strand. This strand initially passes underneath the semicircular loop joining β-strands 2 and 3, and then continues in a path running between these two strands. β-strand 2 runs antiparallel with respect to this new strand, whilst β-strand 3 runs parallel to it. On exit from this β-strand, we next pass through a very long loop region of 18 amino acids. This loop first swings away in an almost perpendicular direction, and then turns back on itself, before stretching off into a new and, at this stage, completely unoccupied region of space. (h) As the long loop ends, our journey takes us through a fourth α-helical region, which is the first long stretch of helix we've encountered so far, comprising 15 amino acids. As we spiral through this helix, we notice that it is cradled within a pocket formed by β-strands 1-4. On the other side of these strands lie the three aforementioned α-helical regions, which are arranged in a triangular formation from this angle, and out beyond these sits the N-terminus of the protein, where the journey began. Upon leaving this helix, we next enter a new loop, shaped like a hump. Once over the top of this hump, we look down to see that the loop ends close to the exposed side of β-strand 3, i.e. where there are no amino acids nearby as yet.

(i) (j)

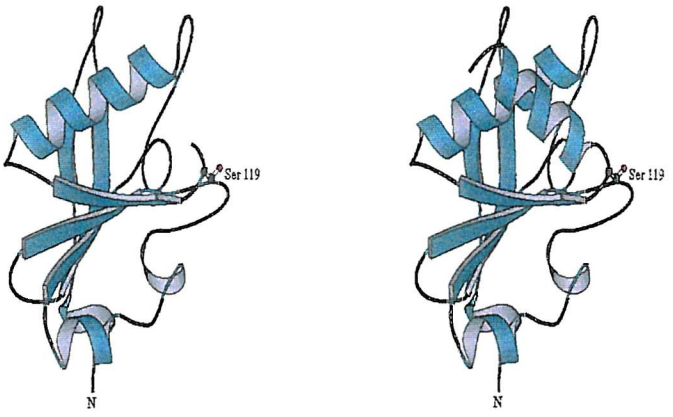


Figure 47. (i) β-strand 5 and putative active site serine. (j) α-helical region 5.

(i) At this point we now find a fifth β-strand which takes up its position next to strand 3, on this, the previously-exposed side of that strand. This new strand runs parallel with strands 3 and 4, and as we travel along it, it not only twists, but also arcs, accentuating the aforementioned β-sheet cradle for α-helix 4.

Upon exit from the fifth β-strand, there is a very short "hump-like" loop, from which a postulated active site serine (serine 119) projects out into space. The short loop terminates at a position partway between the C-terminal end of β-strand 5 and the N-terminal tip of α-helical region 4. (j) This loop then leads into the fifth α-helical section, composed of 12 amino acids. This helix runs partly through the gap in-between β-strand 5 and α-helical region 4, but also projects out of this gap so as to pack almost perpendicularly across the N-terminal third of α-helical region 4. As a consequence of this arrangement, α-helices 4 and 5 are seen to pack together in a way resembling a letter "T".

(k) (l)

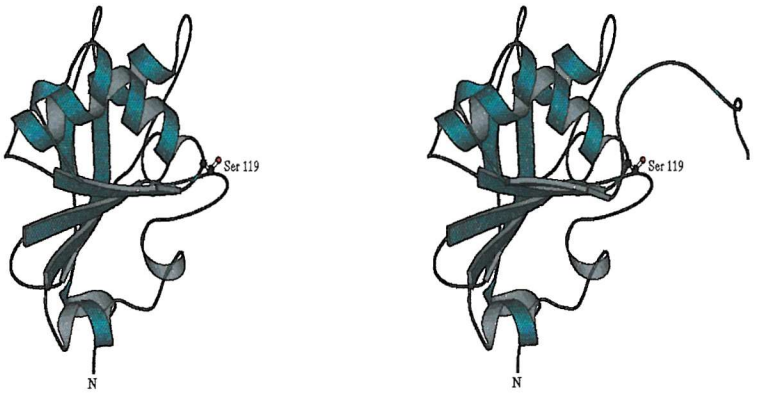


Figure 47. (k) Short bridge-like region. (l) β-strand 6.

(k) Beyond α-helical region 5, there is a short bridge-like structure. This takes the form of a short, straight, 'inclined' stretch of residues building up to a 180° helical twist, and then following this, is an additional short, straight 'declining' section. The end of this bridge-like region projects past the exposed side of β-strand 5, at the latter's N-terminal end.

(l) At this point, a sixth β-strand is found, packing tightly against the exposed side of β-strand 5. Strands 1,3,4,5 and 6 now all run parallel to one another. β-strand 6 comprises 5 amino acids and is relatively flat. Emerging from the β-strand, the protein chain immediately twists off at 90° in a long looping region of 17 amino acids, which twists its way out into a whole new and unoccupied area of space. From the end of this loop, as we look back over the structure which we've explored so far, we can now clearly see a protein divided into three parts. Running through the centre of the protein is the twisting β-sheet, then on one side of this lies the N-terminus and triangular arrangement of small helical coils, whilst on the other side sits the two long helices 4 and 5.

(m) (n) (o)

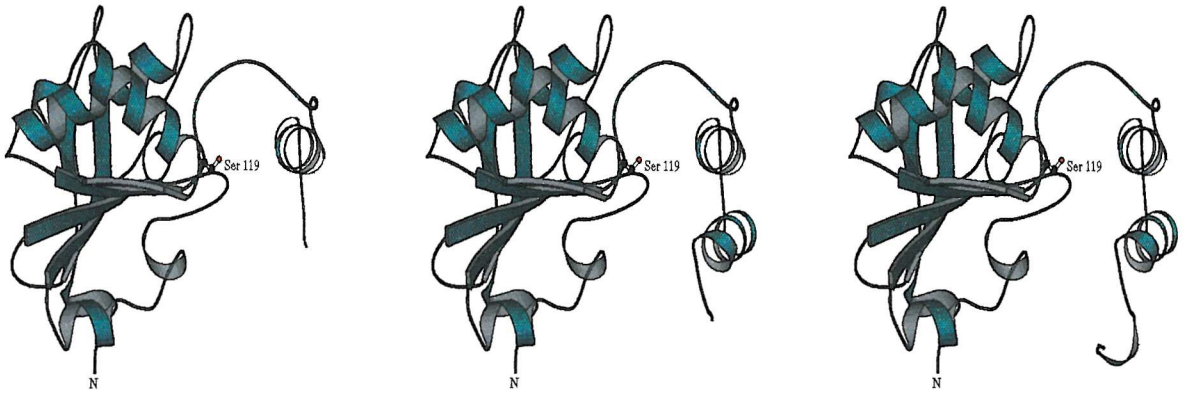


Figure 47. (m) α-helical region 6. (n) α-helical region 7. (o) 180° turn.

(m) The next structural feature we pass through is the sixth α-helical region. This comprises 11 amino acids. This helix passes closest to the tip of the hairpin loop connecting β-strand 3 and α-helical region 2, but still ventures past at a distance. The helix is located in the same 'plane' as the aforementioned β-sheet. At the C-terminal end of the helix, a string of residues projects away at a sharp angle, heading into the half of the structure containing the N-terminus. (n) From this point, a seventh α-helical region of 11 residues emerges, packing up against helix 6 in a parallel arrangement. This helix resides in the half of the structure containing the N-terminus. At the C-terminal end of helical region 7 sits a short bridge-like loop. This loop leaves the helix at a sharp turn, and points off in a direction heading away from α-helical region 6. (o) Next, we are briefly rocked through 180° of helical structure, before travelling along a straight section of 4 amino acids. This straight section runs adjacent to the exposed side of the seventh α-helical region, and also runs more or less parallel with the central axes of α-helical regions 6 and 7.

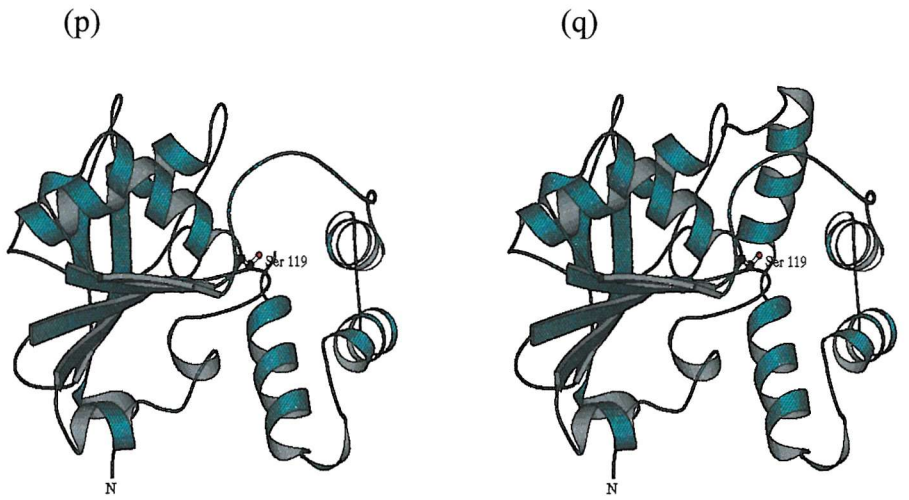


Figure 47. (p) α-helical region 8. (q) α-helical region 9.

(p) Moving on, we next come to α-helix 8, comprising 12 residues. This helix borders the space between helical regions 2 and 7. At its C-terminal end lies a very short arc of structure.

(q) Immediately following this, we see a ninth α-helical region composed of 13 amino acids. This helix is aligned next to the loop region linking β-strand 4 and α-helical region 4. On the opposite side of helix 9 lies the C-terminal tip of helix 6. The sequence protruding from the C-terminal end of helix 9 runs in a perpendicular direction to the helix itself. The path of this sequence is roughly linear, and the whole of this stretch, along with helix 9, lies on the side of the β-sheet containing the large helical regions 4 and 5.

From the end of the linear stretch at the C-terminal end of helix 9, we can now look back once again at how the protein explored so far appears from this vantage point. The protein can now be viewed as being composed of 4 distinct regions. Firstly, there is the β-sheet region. Secondly, there is the T-shaped arrangement of helical regions 4 and 5, which lie on one side of the β-sheet. Thirdly, there is the N-terminus, along with helical regions 1-3, which lie on the other side of the β-sheet. These three distinct areas are all tightly associated together into one region of space. The fourth distinct region evident from this angle is set

(r) (s)

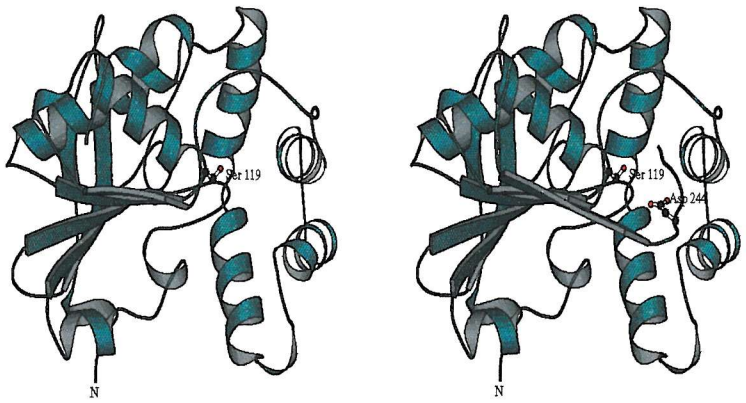


Figure 47. (r) α-helical region 10. (s) β-strand 7 and putative active site aspartate.

aside from the other three areas, lying beyond the C-terminal side of β-strands 3-6. This fourth area, at this stage, comprises α-helical regions 6-9. These four helices are packed together in such a way as to resemble a giant arrow, with helical regions 6 and 7 forming the 'stem of the arrow', and helices 8 and 9 forming the 'arrowhead'.

(r) Continuing on from the linear sequence after α-helical region 9, we next spin through 1½ turns of a new helical region, the tenth encountered so far. This is located nearby the C-terminal end of helical region 5. A short, fairly straight sequence leaves helical region 10, and this terminates very close to the exposed side of β-strand 6, at the N-terminal end.

(s) Before us now lies β-strand 7. This comprises 7 residues, and is reasonably straight, although it does contain a slight kink. It packs adjacent to the exposed side of β-strand 6 and runs parallel with β-strands 1,3,4,5 and 6. The C-terminal loop emerging from this strand first twists towards the postulated active site serine 119, and then curves back away again in a hairpin like fashion. Before the hairpin turn, a second putative active site residue is found, aspartate 244. The aspartate's side chain points inwards, in the general direction of the aforementioned serine.

(t) (u)

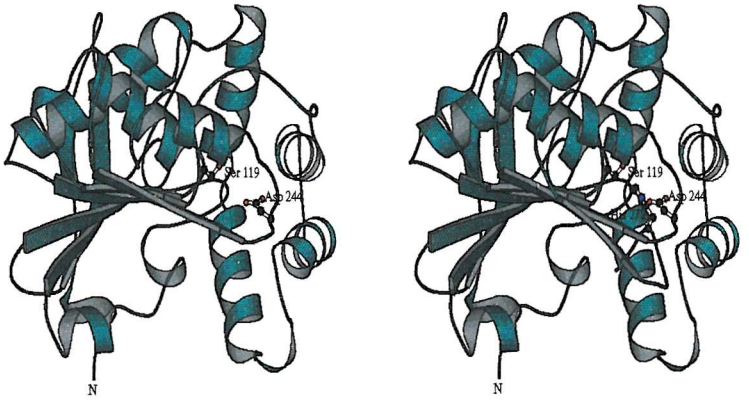


Figure 47. (t) α-helical region 11. (u) β-strand 8 and putative active site histidine.

(t) Moving on, we next rotate through 720° of the eleventh helical region. Composed of 9 residues, this runs partly between α-helical region 10 and β-strand 7. Beyond this helix runs a short arch of sequence, the tip of which lies close to the exposed side of β-strand 7, at the latter's N-terminal end. (u) Next, we pass along a new β-strand, number 8. This is composed of 5 residues and both arcs and twists slightly as we proceed along it. This β-strand packs along the exposed side of β-strand 7, thus extending the β-sheet, and it runs parallel to strands 1,3,4,5,6 and 7. Beyond β-strand 8, the sequence arcs sharply back towards the postulated active site serine 119 once more, before turning into another hairpin bend, and winding away from the serine. At the tip of the hairpin bend sits the third putative active site residue, histidine 272. The side chain of this residue slots in-between serine 119 and aspartate 244, thus completing the putative catalytic triad. The C-terminal tip of the hairpin loop sits on the side of the β-sheet which the protein N-terminus is found on.

(v) (w)

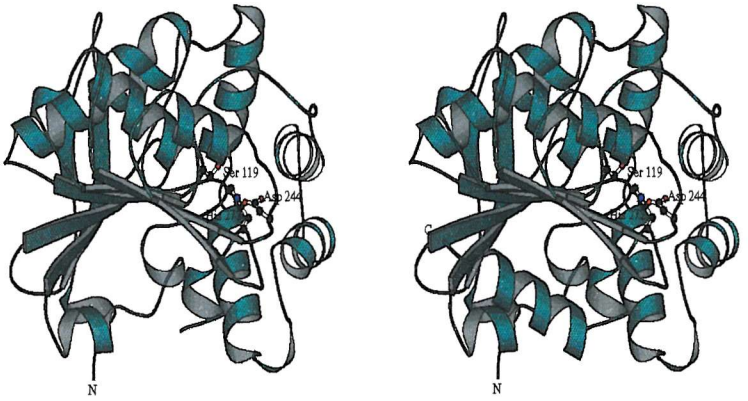


Figure 47. (v) α-helical region 12. (w) α-helical region 13.

(v) We next encounter a ¾ turn of α-helix, which we'll designate α-helical region 12. This packs close to helical regions 2 and 3. These three helical regions together, along with helical region 1, now form a distorted square of helices, with one helical region at each corner. Beyond helical region 12, lies a very short stretch of just 3 residues.

(w) The C-terminus is now in sight. Just preceding this, is one final helix, number 13. This comprises 12 residues and is found in the half of the protein where the N-terminus sits. This helix runs parallel with β-strand 8, and is found cradled in a pocket formed principally by β-strands 3,5,6,7 and 8. The C-terminus then finally rests 4 amino acids beyond this helix.

From the C-terminal position, we can now look back over the entire MhpC protomer structure one final time, to see how all of the pieces of the jigsaw slot together to form the α/β hydrolase fold.

The full MhpC protomer structure may be thought of as five unique segments:

(i) Firstly, there is the central β-sheet, comprising β-strands 1-8. Strands 1,3,4,5,6,7 and 8 run parallel with one another.

(ii) On one side of the β-sheet lies the N-terminus. Also in this region, sandwiched between the N-terminus and the β-sheet are a cluster of 4 small α-helical regions (1,2,3 and 12), arranged in a distorted square. The final feature of this region is the α-helix 13. This sits in a pocket formed by β-strands 3,5,6,7 and 8.

(iii) On the other side of the β-sheet lies four α-helices (4,5,10 and 11). Helices 4 and 5 nestle in a pocket formed by β-strands 1-5. Helix 11 sits atop β-strands 7 and 8, whilst helix 10 packs partway between helices 5 and 11.

(iv) In a separate area of the structure, at the opposite end of the protein to the C-terminus, lies 4 α-helices (6,7,8 and 9). These are arranged in the shape of an arrow. Helices 6 and 7 form the 'stem of the arrow', whilst helices 8 and 9 form the 'arrow head'.

(v) Finally, there is the region between "(i)-(iii)" and "(iv)". This area lacks secondary structural elements, and is composed simply of a number of loop regions. Contained within this region is the postulated active site catalytic triad of serine 119, aspartate 244 and histidine 272.

A combination of these five segments then, defines the MhpC α/β hydrolase fold.

The MhpC active site

As a member of the α/β hydrolase fold family, MhpC possesses a conserved arrangement of putative active site residues, namely, the catalytic triad of Ser 119, Asp 244 and His 272, (see figure 9) which adopt an homologous arrangement in other α/β hydrolase fold enzymes, residing on highly conserved loops [143]. Moreover, mechanistic studies performed with MhpC have indicated the need for an active site nucleophile to participate in phase two of the reaction, where it is likely to be involved in nucleophilic attack of the primary reaction intermediate's carbon 6 (see figure 8). It seems logical that the active site Ser 119 would be able to perform this role via the charge-coupled relay system involving His 272 and Asp 244. Despite this evidence however, it is still not certain what role these residues play in catalysis. Studies involving serine protease inhibitors have shown no inhibition of enzyme activity (Dr. Tim Bugg, personal correspondence), and the active site distance between serine 119 and histidine 272 seems fairly large (3.232Å) from the structure presented here. It is difficult to understand why the serine protease inhibitors are ineffective, assuming the hydrolase functions through its catalytic triad, although the presence of a nearby serine, Ser 49, very close (3.83Å) to the putative Ser 119, could perhaps prevent inhibitor attack due to steric hindrance. Regarding the concerns over the distance between the Ser 119 and His 272, there is a possibility that the presence of substrate near these putative catalytic residues could induce a conformational change in these residues, thus bringing them closer together to allow catalysis.

More evidence supporting the idea that MhpC utilises a catalytic triad in its reactivity, is provided from the MhpC structure determined in this work. Figure 48 below, shows the putative catalytic triad, along with a significant region of electron density which is associated

with the side chain of Ser 119. This observation shows that Ser 119 is capable of binding ligands, and thus is likely to play an active role in the enzyme.

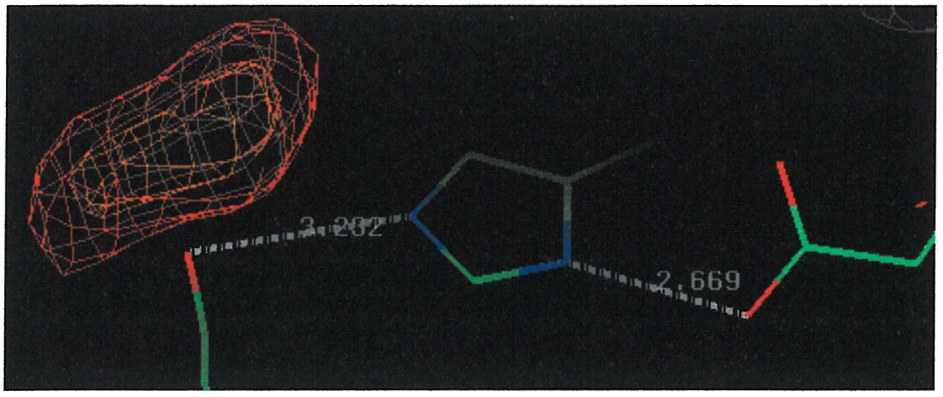


Figure 48. Putative MhpC catalytic triad of serine 119, histidine 272 and aspartate 244. Note the Fo-Fc electron density shown in close proximity to the serine. The density is contoured at both 3σ and 5σ. Image captured from QUANTA [168].

The identity of the ligand shown associated with serine 119 is uncertain. One suggested mechanism of action of MhpC has the serine deprotonating a water molecule, yielding a hydroxyl group, which then nucleophilically attacks the substrate's carbon 6. Thus, one suggestion is that the electron density corresponds to a water molecule, although this is unlikely since the size of the density seems too large.

Another suggestion is that the density represents a metal ion, for example, calcium chloride was used in the mother liquor during crystallisation. However, it again seems unlikely that this is the case, since there is a deficiency in coordination partners in close proximity, which could hold the ion in place.

A third idea is that it is a PMSF derivative. PMSF was used in the preparation of the protein, to help prevent cellular proteases from degrading the hydrolase. However, the density seems too small to be able to account for a PMSF derivative.

A fourth possibility is that it is ethylene glycol, which was present in the mother liquor at 25%. This could resemble the *gem*-diol intermediate resulting from the aforementioned attack of a water molecule on the substrate's carbon 6.

Further experimentation is required in order to determine the identity of this ligand.

A separate intriguing possibility concerning the mechanism of action of MhpC brings into play serine 49 (see figure 49).

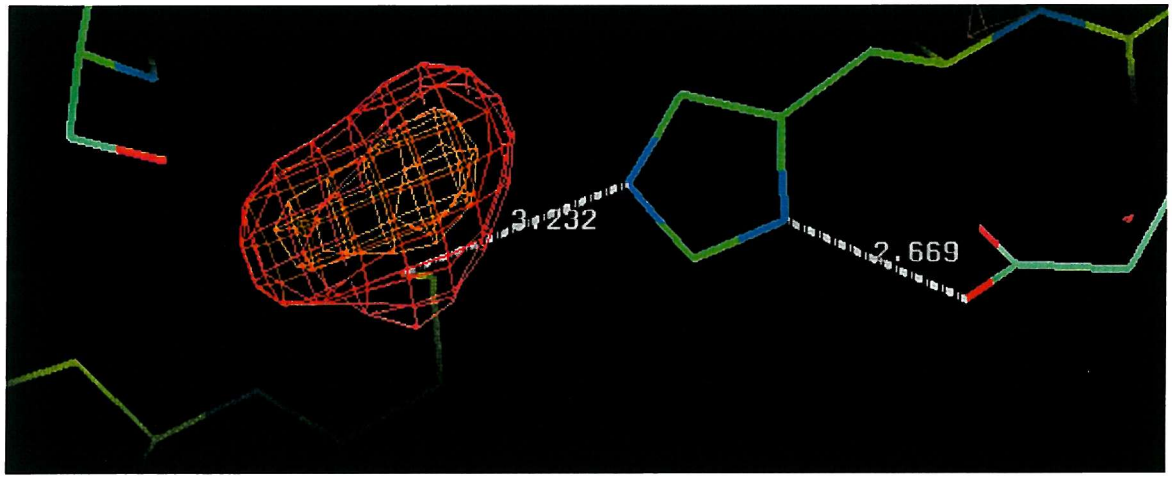


Figure 49. Diagram indicating the possible involvement of Ser 49 (left) in the enzyme mechanism. Electron density shown is contoured at both 3σ and 5σ. Image captured from QUANTA [168].

Assuming Ser 119 nucleophilically attacks the MhpC substrate's carbon 6, this would generate an oxyanion intermediate from the associated ketone group found in the substrate. If this was to be the case, then this oxyanion intermediate would need to be stabilised, in order to favour its formation. This stabilisation could be brought about via hydrogen bonds set up between the oxyanion and adjacent groups. To this end, the Ser 49 side chain might behave as a hydrogen bond donor, along with its main chain amide nitrogen and also the main chain amide nitrogen of the adjacent Gly 50. All of these are arranged nearby the electron density found at Ser 119, and may form what is known as an oxyanion hole [189]. An oxyanion hole

is an arrangement of at least two main chain amide groups, and possibly other determinants, which form a binding site whose role it is to stabilise the reaction transition state via hydrogen bonding [189]. Several enzymes have been shown to possess an oxyanion hole at their active sites [190,191,192], and at least one hydrolytic enzyme, cutinase, from the fungus *Fusarium solani pisi*, has not only been shown to possess an oxyanion hole [189], but also has an α/β hydrolase fold structure [193], as with MhpC, thus highlighting the real possibility that serine 49 may be involved in the mechanism.

The most recent mechanistic work has shown no evidence for an acyl-enzyme intermediate during the MhpC reaction [141]. This means that the serine probably acts as a base to allow the attack of water at the substrate's carbon-6 during the second phase of the reaction. This in turn would generate the aforementioned *gem*-diol intermediate. If this is the case, then clearly the oxyanion hole would not play a part in the stabilisation of this uncharged intermediate.

Another interesting feature in the structure of MhpC is a region of strong density located a short distance away from the putative catalytic triad. This region is shown in figure 50 below.

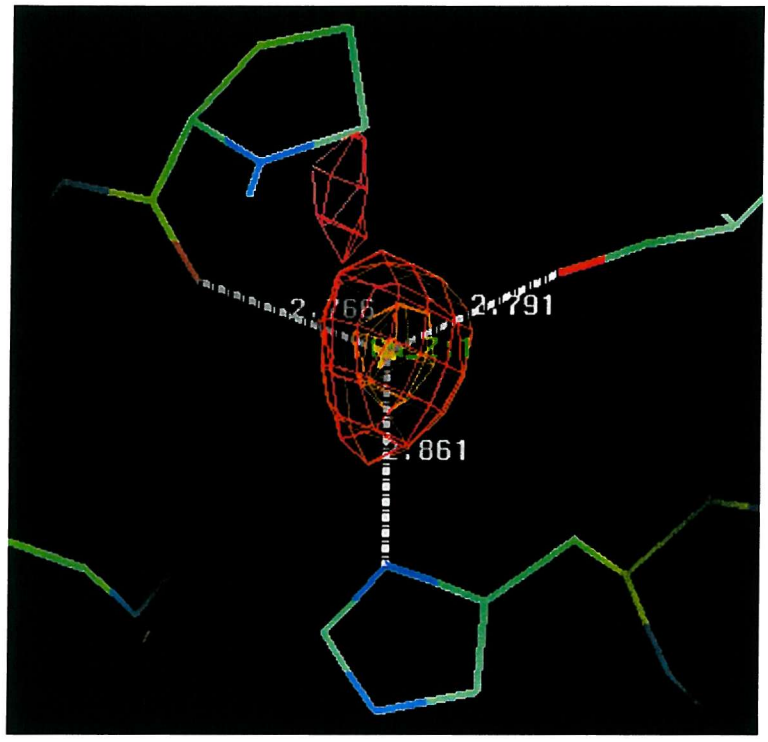


Figure 50. Additional region of strong electron density seen in the MhpC Fo-Fc map. The contour levels are set at 3σ and 5σ. Hydrogen bonding appears possible to the surrounding groups, Ser 94, His 123 and the carbonyl oxygen of Pro 223. Image captured from QUANTA [168].

From this picture, the bond lengths shown imply that the density corresponds to a hydrogen bonded water molecule in the structure. The density appears relatively strong, and thus the water is likely to have a low temperature factor associated with it, meaning it is a very well ordered molecule. A water molecule in this location of the structure could be used to tether the coordinated His 123 in a position from which it can bind to the substrate's carbon 9 carboxyl group, thus helping to fix the substrate in place in preparation for catalysis. Further evidence to support this hypothesis comes from a comparison of the region between the putative active site and this area of density in MhpC and in 2-hydroxy-6-oxo-6-phenylhexa-

2,4-dienoate hydrolase (BphD), an α/β hydrolase fold enzyme from *Rhodococcus Sp.*, which is homologous to MhpC and which catalyses a similar reaction. Figure 51 shows a comparison of the layout of the two structures in this region.

(a) (b)

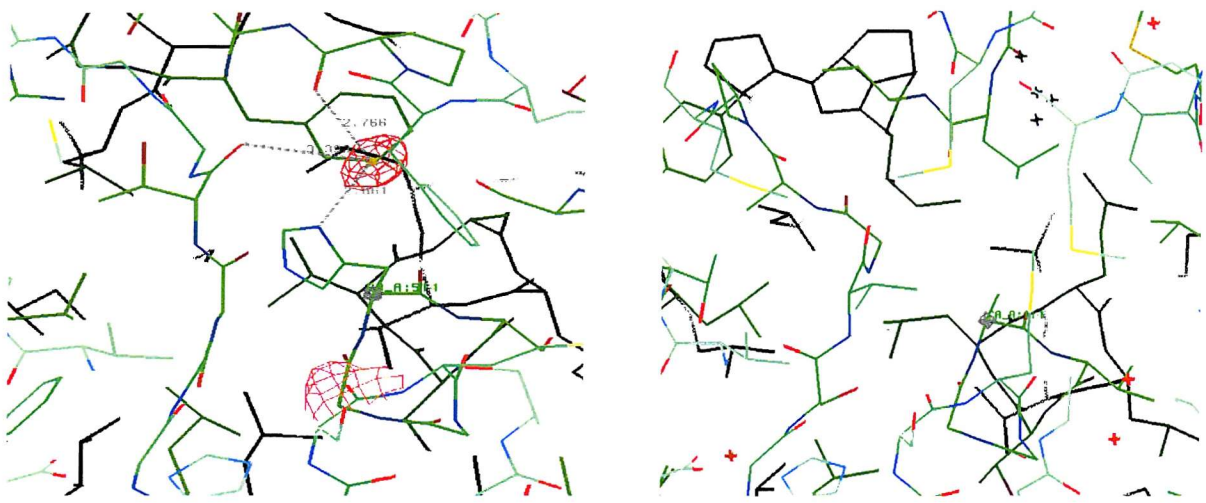


Figure 51. Comparison of a possible substrate binding region in (a) MhpC and (b) BphD. (a) In MhpC, His 123 lies in an area found between the putative active site residue Ser 119 and the unknown extra region of density shown in the top half of the picture. Ser 119 is partially surrounded by electron density from the undetermined ligand as discussed in the main text, this is to help locate its position. The contour level used to display the electron density was 4σ. (b) In BphD, there is no equivalent histidine to the MhpC His 123, nor is there a hydrophilic pocket present in this region. Instead there exists a hydrophobic pocket not present in the MhpC structure, and the MhpC His 123 has been replaced by Ala 114. The BphD coordinates were obtained from the Protein Data Bank [37], and the structure was solved by Nandhagopal, *et al* [194]. Pictures captured from QUANTA [168].

From figure 51, it can be seen that where MhpC has a relatively hydrophilic region near the putative active site, centred around His 123, the equivalent region in the BphD structure is very hydrophobic. In the MhpC structure, the hydrophilic pocket includes residues Ser 94, His 123, the backbone carbonyl group of Thr 146 and the backbone carbonyl group of Pro 223. At the same time, the side chains of Met 120 and Phe 222 project away from this pocket.

In BphD, however, the pocket is strongly hydrophobic, comprising residues Val 85, Met 111, Ala 114, Leu 117, Leu 213 and Ile 215.

This is a significant observation, since the only difference between the substrates of the two hydrolases is that the MhpC substrate has a propionate group attached to carbon 6, whilst the BphD substrate has an aromatic phenyl group in an equivalent position. Thus, it is proposed that the pockets described here, and shown in figure 51, represent the binding sites for the substrate's propionate group in the MhpC structure, and the substrate's phenyl group in the BphD structure.

One final point of interest from the structure is how the subunits are associated in the native MhpC dimer. Figure 52 shows how three of the four MhpC protomers present in the asymmetric unit, pack together in the crystal. From this figure, it can be seen that in the crystal, the various MhpC protomers associate with each other through three different interfaces. Of these, the most stable inter-protomer interaction appears to be that involving a β-strand from each protomer. Where this occurs between two protomers, the result is a combination of their individual β-sheets, effectively forming, therefore, a 16 β-strand β-sheet, which would be predicted to be a favourable structural feature of a dimeric MhpC enzyme. In support of this, the other two inter-protomer interactions seen in the asymmetric unit of the MhpC crystal are unlikely to be stable interactions in solution, since one of these regions is a poor helix-helical turn association, whilst the other is an interaction between two flexible loop regions, which could associate together in a relatively stable way in the more rigid environment of the crystal, but which would be unlikely to form a stable association in solution.

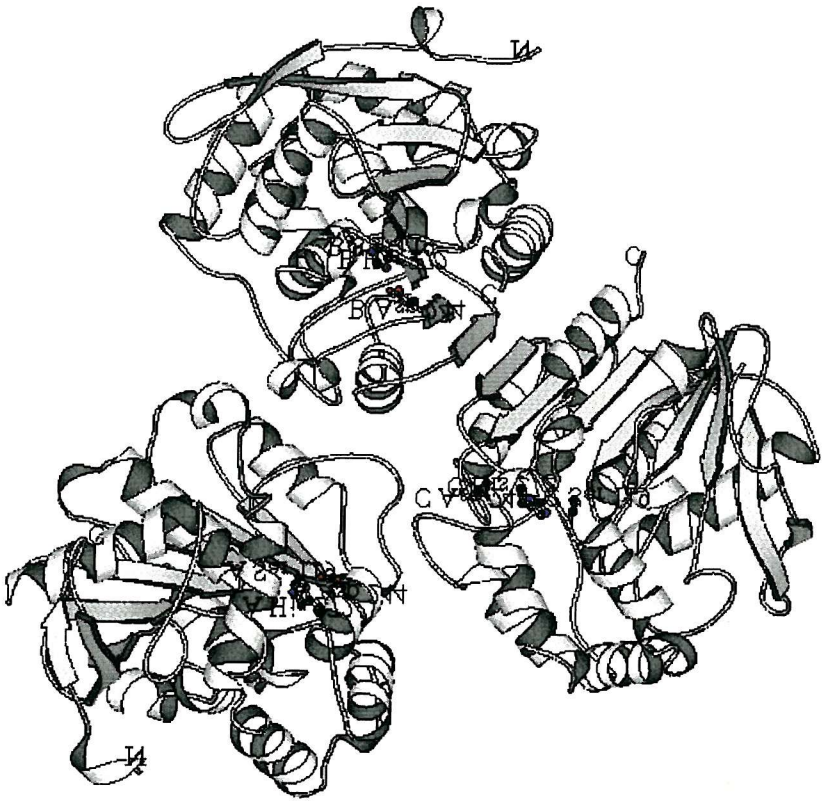


Figure 52. Picture showing how three of the four MhpC molecules in the asymmetric unit associate in the crystal. From this view, an idea can be gained as to how the native dimeric protein's subunits are associated. The N and C termini are labelled, as are the putative active site residues. Picture generated by MOLSCRIPT [35].

Appendix

6. Appendix

Appendix 1. The events responsible for crystal growth in a vapour diffusion experiment

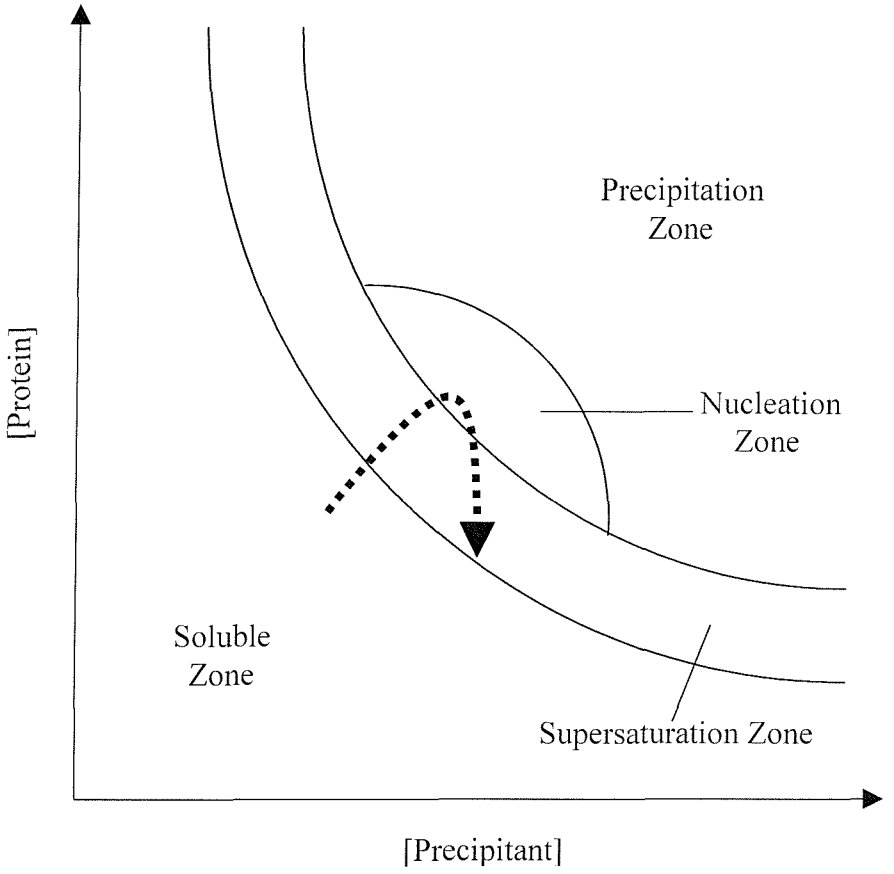


Figure 53. A typical protein solubility curve summarising the events which occur during the course of a hanging-drop vapour diffusion crystallisation. The figure is divided into four zones: (1) the soluble zone, where protein is freely soluble in solution, (2) the supersaturated zone, where protein approaches the limits of its solubility, (3) the nucleation zone, where individual protein molecules come out of solution and serve as origins from which crystal growth may be propagated, and (4) the precipitation zone, where the majority of the protein becomes insoluble and aggregates together to produce an amorphous precipitate. The dashed line indicates how the protein and precipitant concentrations in the hanging drop vary with time, on a scale of days to weeks, culminating at the tip of the arrowhead, at which point the crystals are fully grown.

During a hanging drop-vapour diffusion experiment, crystallisation is achieved via a multistep process. This is summarised in figure 53. The dashed arrow in the figure shows how the conditions in the drop change with time. Following the arrow, it can be seen that the preliminary changes in the drop involve a simple concentration of the protein and precipitant levels. As this happens, the protein approaches the limits of its solubility (as indicated by the

supersaturation zone in the figure), and eventually, a very small number of protein molecules become insoluble (equivalent to the arrow entering the nucleation zone in the figure). Once insoluble, these protein molecules act as nuclei, or starting points, upon which the crystals can now be assembled. After this process has begun, crystal assembly continues steadily, since additional protein molecules are still in a state of supersaturation, and thus eagerly accept the opportunity to come out of solution and join the stable crystal growth. As protein falls out of solution in this way, then the protein concentration in the drop decreases (as shown in the figure), until, after several days or weeks, the protein level is sufficiently low so as to no longer be in a supersaturated state (equivalent to the point reached by the arrowhead in the figure), at which point crystal growth ceases.

References

7. References

[1] Wheatley, L.M. & Platts-Mills T.A.E. (1996). Perennial allergens and the asthma epidemic. *Science & Medicine* **May/June**, 6-13

[2] Deleuran, M., Ellingsen, A.R., Paludan, K., Schou, C. & ThestrupPedersen, K. (1998). Purified *Der p*1 and *p*2 patch tests in patients with atopic dermatitis: Evidence for both allergenicity and proteolytic irritancy. *Acta Dermato-Venereologica* **78**, No.4, 241-243

[3] Tan, B.B., Weald, D., Strickland, I. & Friedmann, P.S. (1996). Double-blind controlled trial of effect of house dust mite allergen avoidance on atopic dermatitis. *Lancet* **347**, No.8993, 15-18

[4] Ichikawa, S., Hatanaka, H., Yuuki, T., Iwamoto, N., Kojima, S., Nishiyama, C., Ogura, K., Okumura, Y. & Inagaki, F. (1998). Solution structure of *Der f*2, the major mite allergen for atopic diseases. *J. Biol. Chem.* **273**, No.1, 356-360

[5] Chapman, M.D., Rowntree, S., Mitchell, E.B., Defuenmajor, M.C.D. & Platts-Mills, T.A.E. (1983). Quantitative assessments of IgG and IgE antibodies to inhalant allergens in patients with atopic dermatitis. *J. Allergy & Clin. Immunol.* **72**, 27-33

[6] Krilis, S., Baldo, B.A., Sutton, R. & Basten, A. (1984). Antigens and allergens from the common house dust mite *Dermatophagoides pteronyssinus* 1. Demonstration of multiple allergens by immunochemical and biologic analyses. *J. Allergy Clin. Immunol.* **74**, 132-141

[7] Floyer, J. (1698) *A Treatise of the Asthma.*, R. Wilkin, London.

[8] Kern, R.A. (1921) Dust sensitisation in bronchial asthma. *Med. Clin. North Am.* **5**; 751-758

[9] Spivacke, C.A. & Grove, E.F. (1925). Studies in hypersensitiveness. XIV: A study of the house dust atopen in asthma. *J. Immunol.* **10**, 465-470

[10] Voorhorst, R., Spieksma-Boezeman, M.I.A. & Spieksma, F.Th.M. (1964). Is a mite (*Dermatophagoides sp.*) the producer of the house dust allergen? *Allergie und Asthma.* **10**, 329-334

[11] Voorhorst, R., Spieksma, F.Th.M., Varekamp, H., Leupen, M.J. & Lyklema, A.W. (1967). The house dust mite (*Dermatophagoides pteronyssinus*) and the allergens it produces. Identity with the house dust allergen. *J. Allergy. Clin. Immunol.* **39**, 325-335

[12] Chapman, M.D. & Platts-Mills, T.A.E. (1980). Purification and characterisation of the major allergen from *Dermatophagoides pteronyssinus* - antigen *p*1. *J. Immunol.* **125**, No.2, 587-592

[13] Tovey, E.R., Chapman, M.D. & Platts-Mills, T.A.E. (1981). Mite faeces are a major source of house dust allergens. *Nature* **289**, No.5798, 592-593

[14] Chua K.Y., Stewart, G.A., Thomas, W.R., Simpson, R.J., Dilworth, R.J., Plozza, T.M. & Turner, K.J. (1988). Sequence-analysis of cDNA coding for a major house dust mite allergen, *Der p*1. Homology with cysteine proteases. *J.Exp. Med.* **167**, 175-182

[15] Dilworth, R.J., Chua, K.Y. & Thomas, W.R. (1991). Sequence analysis of cDNA coding for a major house dust mite allergen, *Der f*1. *Clin. & Exp. Allergy* **21**, 25-32

[16] Herbert, C.A., Holgate, S.T., Robinson, C., Thompson, P.J. & Stewart, G.A. (1990). Effect of mite allergen on permeability of bronchial mucosa. *The Lancet* **336**, No.8723, 1132

[17] Hewitt, C.R.A., Brown, A.P., Hart, B.J. & Pritchard, D.I. (1995). A major house dust mite allergen disrupts the immunoglobulin-E network by selectively cleaving CD23 - innate protection by antiproteases *J. Exp. Med.* **182**, 1537-1544

[18] Attwood, T.K., Payne, A.W.R., Michie, A.D. & Parry-Smith, D.J. (1997). A Colour INteractive Editor for Multiple Alignments – CINEMA. *EMBnet news*, **3** (3)

[19] Chua, K.Y., Kehal, P.K. & Thomas, W.R. (1993). Sequence polymorphisms of cDNA clones encoding the mite allergen *Der p*1. *Int. Arch. Allergy Immunol.* **101**, 364-368

[20] Johnson, M.S. (1990). Modelling and Biocomputing group, Center for Biotechnology, P.O. Box 123, FIN-20521 Turku, Finland. MALIGN: ©1990 Kramsku, Finland

[21] Cohen, L.W., Coghlan, V.M. & Dihel, L.C. (1986). Cloning and sequencing of papain-encoding cDNA. *Gene* **48**, 219-227

[22] Podivinsky, E., Forster, R.L.S. & Gardner, R.C. (1989). Nucleotide sequence of actinidin, a kiwi fruit protease. *Nucleic Acids Res.* **17**, 8363-8363

[23] Guenette, R.S., Mooibroek, M., Wong, K., Wong, P. & Tenniswood, M. (1994). Cathepsin B, a cysteine protease implicated in metastatic progression, is also expressed during regression of the rat prostate and mammary glands. *Eur. J. Biochem.* **226**, 311-321

[24] Ishidoh, K., Imajoh, S., Emori, Y., Ohno, S., Kawasaki, H., Minami, Y., Kominami, E., Katunuma, N. & Suzuki, K. (1987). Molecular cloning and sequencing of cDNA for rat cathepsin H. Homology in pro-peptide regions of cysteine proteinases. *FEBS Letts.* **226**, 33-37

[25] Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45-48. http://expasy.cbr.nrc.ca/sprot/

[26] Kalsheker, N.A., Deam, S., Chambers, L., Sreedharan, S., Brocklehurst, K. & Lomas, D.A. (1996). The house dust mite allergen *Der p*1 catalytically inactivates alpha(1)-antitrypsin by specific reactive centre loop cleavage: A mechanism that promotes airway inflammation and asthma. *Biochem. Biophys. Res. Comm.* **221**, 59-61

[27] Corris, P.A. & Dark, J.H. (1993). Etiology of asthma: lessons from lung transplantation. *Lancet* **341**, No.8857, 1369-1371

[28] Mori, L., Kleimberg, J., Mancini, C., Bellini, A., Marini, M. & Mattoli, S. (1995). Bronchial epithelial cells of atopic patients with asthma lack the ability to inactivate allergens. *Biochem. Biophys. Res. Comm.* **217**, No.3, 817-824

[29] Wang, J.Y., Kishore, U., Lim, B.L., Strong, P. & Reid, K.B.M. (1996). Interaction of human lung surfactant proteins A and D with mite (*Dermatophagoides pteronyssinus*) allergens. *Clin. Exp. Immunol.* **106**, No.2, 367-373

[30] Caulfield, J.J., Hawrylowicz, C.M., Kemeny, D.M. & Lee, T.H. (1997). GM-CSF increases the ability of cultured macrophages to support autologous CD4(+) T-cell proliferation in response to *Dermatophagoides pteronyssinus* and PPD antigen. *Immunology* **92**, No.1, 123-130

[31] Stacey, M.A., Sun, G., Vassalli, G., Marini, M., Bellini, A. & Mattoli, S. (1997). The allergen *Der p*1 induces NF-kappa B activation through interference with I kappa B alpha function in asthmatic bronchial epithelial cells. *Biochem. Biophys. Res. Comm.* **236**, No.2, 522-526

[32] Comoy, E.E., Pestel, J., Duez, C., Stewart, G.A., Vendeville, C., Fournier, C., Finkelman, F., Capron, A. & Thyphronitis, G. (1998). The house dust mite allergen, *Dermatophagoides pteronyssinus*, promotes type 2 responses by modulating the balance between IL-4 and IFN-gamma. *J. Immunol.* **160**, No.5, 2456-2462

[33] Topham, C.M., Srinivasan, N., Thorpe, C.J., Overington, J.P. & Kalsheker, N.A. (1994). Comparative modeling of major house dust mite allergen *Der p*1 - structure validation using an extended environmental amino-acid propensity table. *Protein Eng.* **7**, 869-894

[34] Evans, S.V. (1993). SETOR - Hardware-lighted three-dimensional solid model representations of macromolecules. *J. Mol. Graph.* **11**, No.2, 134-138

[35] Kraulis, P.J. (1991). MOLSCRIPT - A program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* **24**, No.Pt5, 946-950

[36] Pickersgill, R.W., Harris, G.W. & Garman, E. (1992). Structure of monoclinic papain at 1.60Å resolution. *Acta Cryst. B* **48**, No.Pt1, 59-67

[37] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids Research*, **28**, 235-242. http://www.rcsb.org/pdb/

[38] Collins, S.P., Ball, G., Vonarx, E., Hosking, C., Shelton, M., Hill, D. & Howden, M.E.H. (1996). Absence of continuous epitopes in the house dust mite major allergens *Der p*1 from *Dermatophagoides pteronyssinus* and *Der f*1 from *Dermatophagoides farinae. Clin. Exp. Allergy* **26**, No.1, 36-42

[39] Conradi, H. (1903). Ueber löslishe, durch aseptische autolyse, erhaltene giftstoffe von ruhr- und typhus bazillen. *Dtsch. Med. Wochenschr.* **29**, 26-28

[40] Thompson, M.R., Steinberg, M.S., Gemski, P., Formal, S.B. & Doctor, B.P. (1976). Inhibition of in vitro protein synthesis by *Shigella dysenteriae* 1 toxin. *Biochem. Biophys. Res. Comm.* **71**, 783-788

[41] O'Brien, A.D., Thompson, M.R., Cantey, J.R. & Formal, S.B. (1977). Production of a *Shigella dysenteriae*-like toxin by pathogenic *Escherichia coli*. *Abstr. Annu. Meet. Am. Soc. Microbiol.* **B103**, 32

[42] Scotland, S.M., Smith, H.R. & Rowe, B. (1985). Two distinct toxins active on vero cells from *Escherichia coli* O157. *Lancet*, **2**, No. 8460, 885-886

[43] Yao, Q.Z., Yu, M.M., Ooi, L.S.M., Ng, T.B. Chang, S.T., Sun, S.S.M. & Ooi, V.E.C. (1998). Isolation and characterisation of type 1 ribosome-inactivating protein from fruiting bodies of the edible mushroom (Volvariella volvacea). *J. Agric. & Food Chem.* **46**, No. 2, 788-792

[44] Jimenez, A. & Vasquez, D. (1985). Plant and fungal protein and glycoprotein toxins inhibiting eukaryote protein synthesis. *Ann. Rev. Microbiol.* **39**, 649-672

[45] Stirpe, F. & Barbieri, L. (1986). Ribosome-inactivating proteins up-to-date. *FEBS Letts.* **195**, 1-8

[46] Fong, W.P., Wong, R.N.S., Go, T.T.M. & Yeung, H.W. (1991). Enzymatic-properties of ribosome-inactivating proteins (RIPs) and related toxins. *Life Sciences* **49**, 1859-1869

[47] Ren, J.S., Wang, Y.P., Dong, Y.C. & Stuart, D.I. (1994). The *N*-glycosidase mechanism of ribosome-inactivating proteins implied by crystal-structures of alpha-momorcharin. *Structure* **2**, No.1, 7-16

[48] Pu, Z., Lu, B.-Y., Liu, W.-Y. & Jin, S.-W. (1996). Characterisation of the enzymatic mechanism of γ-momorcharin, a novel ribosome-inactivating protein with lower molecular weight of 11,500 purified from the seeds of bitter gourd (*Momordica charantia). Biochem. Biophys. Res. Comm.* **229**, 287-294

[49] Xiong, C.Y. & Zhang, Z.C. (1998). Isolation, purification and characterisation of a group of novel small molecular ribosome inactivating proteins - LuffinS from the seeds of *Luffa cylindrica. Acta Biochim. Biophys. Sinica* **30**, No. 2, 142-146

[50] Walsh, T.A., Morgan, A.E. & Hey, T.D. (1991). Characterisation and molecular cloning of a proenzyme form of a ribosome-inactivating protein from maize. *J. Biol. Chem.* **266**, 23422-23427

[51] O'Brien, A.D. & Holmes, R.K. (1987). Shiga and Shiga-like toxins. *Microbiol. Reviews* **51**, No. 2, 206-220

[52] Hartley, M.R., Chaddock, J.A. & Bonness, M.S. (1996). The structure and function of ribosome-inactivating proteins. *Trends in Plant Sci.* **1**, No. 8, 254-260

[53] Olsnes, S., Fernandez-Puentes, C., Carrasco, L. & Vazquez, D. (1975). Ribosome inactivation by the toxic lectins abrin and ricin – Kinetics of the enzymic activity of the toxin A-chains. *Eur. J. Biochem.* **60**, 281-288

[54] Lord, J.M., Roberts, L.M. & Robertus, J.D. (1994). Ricin: structure, mode of action, and some current applications. *FASEB Journal* **8**, 201-208

[55] Richardson, P.T., Westby, M., Roberts, L.M., Gould, J.H., Colman, A. & Lord, J.M. (1989). Recombinant proricin binds galactose but does not depurinate 28-S ribosomal-RNA *FEBS Letts.* **255**, No.1, 15-20

[56] Roberts, L.M. & Lord, J.M. (1981). The synthesis of ricinus-communis agglutinin - cotranslational and posttranslational modification of agglutinin polypeptides. *Eur. J. Biochem.* **119**, No.1, 31-41

[57] Lord, J.M. (1985a). Synthesis and intracellular-transport of lectin and storage protein precursors in endosperm from castor bean. *Eur. J. Biochem.* **146**, No.2, 403-409

[58] Lord, J.M. (1985b). Precursors of ricin and ricinus-communis agglutinin - glycosylation and processing during synthesis and intracellular-transport. *Eur. J. Biochem.* **146**, No.2, 411-416

[59] Frigerio, L., Vitale, A., Lord, J.M., Ceriotti, A., Roberts, L.M. (1998). Free ricin A-chain, proricin, and native toxin have different cellular fates when expressed in tobacco protoplasts. *J. Biol. Chem.* **273**, 14194-14199

[60] Hiraiwa, N., Kondo, M., Nishimura, M. & HaraNishimura, I. (1997). An aspartic endopeptidase is involved in the breakdown of propeptides of storage proteins in protein-storage vacuoles of plants. *Eur. J. Biochem.* **246**, No.1, 133-141

[61] Hara-Nishimura, I., Inoue, K. & Nishimura, M. (1991). A unique vacuolar processing enzyme responsible for conversion of several proprotein precursors into the mature forms. *FEBS Letts.* **294**, 89-93

[62] Frigerio, L. & Roberts, L.M. (1998). The enemy within: ricin and plant cells. *J. Exp. Bot.* **49**, No. 326, 1473-1480

[63] Legname, G., Bellosta, P., Gromo, G., Modena, D., Keen, J.N., Roberts, L.M. & Lord, J.M. (1991). Nucleotide-sequence of cDNA coding for dianthin-30, a ribosome-inactivating protein from *Dianthus caryophyllus*. *Biochim. Biophys. Acta* **1090**, No.1, 119-122

[64] Carzaniga, R., Sinclair, L., Fordhamskelton, A.P., Harris, N. & Croy, R.R.D. (1994). Cellular and subcellular-distribution of saporins, type-1 ribosome-inactivating proteins, in soapwort (*saponaria-officinalis* l). *Planta* **194**, No. 4, 461-470

[65] Ready, M.P., Brown, D.T. & Robertus, J.D. (1986). Extracellular localisation of pokeweed antiviral protein. *Proc. Natl. Acad. Sci. USA* **83**, No.14, 5053-5056

[66] Leah, R., Tommerup, H., Svendsen, I. & Mundy, J. (1991). Biochemical and molecular characterisation of three barley seed proteins. *J. Biol. Chem.* **266**, 1564-1573

[67] Bass, H.W., Webster, C., Obrian, G.R., Roberts, J.K.M., Boston, R.S. (1992). A maize ribosome-inactivating protein is controlled by the transcriptional activator opaque-2. *Plant Cell* **4**, No.2, 225-234

[68] Taylor, B.E. & Irvin, J.D. (1990). Depurination of plant ribosomes by pokeweed antiviral protein. *FEBS Letts.* **273**, No.1-2, 144-146

[69] Chaudhry, B., Mulleruri, F., Cameronmills, V., Gough, S., Simpson, D., Skriver, K. & Mundy, J. (1994). The barley 60kDa jasmonate-induced protein (Jip60) is a novel ribosome-inactivating protein. *Plant Journal* **6**, No.6, 815-824

[70] Simpson, J.C., Smith, D.C., Roberts, L.M., Lord, J.M. (1998). Expression of mutant dynamin protects cells against diphtheria toxin but not against ricin. *Exptl. Cell Res.* **239**, No.2, 293-300

[71] Van Deurs, B., Sandvig, K., Petersen, O.W., Olsnes, S., Simons, K. & Griffiths, G. (1988). Estimation of the amount of internalised ricin that reaches the trans-Golgi network. *J. Cell Biol.* **106**, No.2, 253-267

[72] Rapak, A., Falnes, P.O. & Olsnes, S. (1997). Retrograde transport of mutant ricin to the endoplasmic reticulum with subsequent translocation to cytosol. *Proc. Natl. Acad. Sci. USA* **94**, No.8, 3783-3788

[73] Lord, J.M. & Roberts, L.M. (1998). Toxin entry: Retrograde transport through the secretory pathway. *J. Cell Biol.* **140**, No.4, 733-736

[74] Wiertz, E.J.H.J., Jones, T.R., Sun, L., Bogyo, M., Geuze, H.J. & Ploegh, H.L. (1996). The human cytomegalovirus US11 gene product dislocates MHC class I heavy chains from the endoplasmic reticulum to the cytosol. *Cell*, **84**, No.5, 769-779

[75] Hiller, M.M., Finger, A., Schweiger, M. & Wolf, D.H. (1996). ER degradation of a misfolded luminal protein by the cytosolic ubiquitin-proteasome pathway. *Science* **273**, No.5282, 1725-1728

[76] Wiertz, E.J.H.J., Tortorella, D., Bogyo, M., Yu, J., Mothes, W., Jones, T.R., Rapoport, T.A., Ploegh, H.L. (1996). Sec61-mediated transfer of a membrane protein from the endoplasmic reticulum to the proteasome for destruction. *Nature* **384**, No.6608, 432-438

[77] Plemper, R.K., Bohmler, S., Bordallo, J., Sommer, T., Wolf, D.H. (1997). Mutant analysis links the translocon and BiP to retrograde protein transport for ER degradation. *Nature* **388**, No.6645, 891-895

[78] Pilon, M., Schekman, R. & Romisch, K. (1997). Sec61p mediates export of a misfolded secretory protein from the endoplasmic reticulum to the cytosol for degradation. *EMBO J.* **16**, No.15, 4540-4548

[79] Hazes, B. & Read, R.J. (1997). Accumulating evidence suggests that several AB-toxins subvert the endoplasmic reticulum-associated protein degradation pathway to enter target cells. *Biochemistry* **36**, No.37, 11051-11054

[80] Walker, D., Chaddock, A.M., Chaddock, J.A., Roberts, L.M., Lord, J.M. & Robinson, C. (1996). Ricin A-chain fused to a chloroplast-targeting signal is unfolded on the chloroplast surface prior to import across the envelope membranes. *J. Biol. Chem.* **271**, 4082-4085

[81] Pohl, P., Antonenko, Y.N., Evtodienko, V.Y., Pohl, E.E., Saparov, S.M., Agapov, I.I. & Tonevitsky, A.G. (1998). Membrane fusion mediated by ricin and viscumin. *Biochim. Biophys. Acta-Biomembranes* **1371**, No.1, 11-16

[82] Walzel, H., Bremer, H. & Gabius, H.J. (1993) in *Lectins and Glycobiology* (Gabius, H.J. & Gabius, S., eds.), pp. 356-361, Springer, Berlin

[83] Siegel, D.P., Banschbach, J., Alford, D., Ellens, H., Lis, L.J., Quinn, P.J., Yeagle, P.L. & Bentz, J. (1989). Physiological levels of diacylglycerols in phospholipid-membranes induce membrane-fusion and stabilize inverted phases. *Biochemistry* **28**, No.9, 3703-3709

[84] Hudson, T.H. & Grillo, F.G. (1991). Brefeldin-A enhancement of ricin A-chain immunotoxins and blockade of intact ricin, modeccin, and abrin. *J. Biol. Chem.* **266**, No.28, 18586-18592

[85] Bau, M.Y. & Draper, R.K. (1993). Ricin intoxicates END4 mutants that have an aberrant Golgi-complex. *J. Biol. Chem.* **268**, 19939-19942

[86] Okimoto, T., Seguchi, T., Ono, M., Nakayama, Y., Funatsu, G., Fujiwara, T., Ikehara, Y. & Kuwano, M. (1993). Brefeldin-A protects ricin-induced cytotoxicity in human cancer KB cell-line, but not in its resistant counterpart with altered Golgi structures. *Cell Struct. Funct.* **18**, No.4, 241-251

[87] Dowd, P.F., Mehta, A.D. & Boston, R.S. (1998). Relative toxicity of the maize endosperm ribosome-inactivating protein to insects. *J. Agric. & Food Chem.* **46**, No. 9, 3775-3779

[88] Stirpe, F., Barbieri, L., Battelli, M.G., Soria, M. & Lappi, D.A. (1992). Ribosome-inactivating proteins from plants - present status and future-prospects. *Biotechnology* **10**, 405-412

[89] Vitetta, E.S., Krolick, K.A., Miyamainaba, M., Cushley, W. & Uhr, J.W. (1983). Immunotoxins - A new approach to cancer-therapy. *Science* **219**, 644-650

[90] McGrath, M.S., Hwang, K.M., Caldwell, S.E., Gaston, I., Luk, K.C., Wu, P., Ng, V.L., Crowe, S., Daniels, J., Marsh, J., Deinhart, T., Lekas, P.V., Vennari, J.C., Yeung, H.W. & Lifson, J.D. (1989). GLQ223 - an inhibitor of human immunodeficiency virus-replication in acutely and chronically infected-cells of lymphocyte and mononuclear phagocyte lineage. *Proc. Natl. Acad. Sci. USA* **86**, 2844-2848

[91] Zarling, J.M., Moran, P.A., Haffar, O., Sias, J., Richman, D.D., Spina, C.A., Myers, D.E., Kuebelbeck, V., Ledbetter, J.A. & Uckun, F.M. (1990). Inhibition of HIV replication by

pokeweed antiviral protein targeted to CD4+ cells by monoclonal-antibodies. *Nature* **347**, 92-95

[92] Marcil, J., Ravindranath, N. & Sairam, M.R. (1993). Cytotoxic activity of lutropin-gelonin conjugate in mouse Leydig tumour cells - potentiation of the hormonotoxin activity by different drugs. *Molec. & Cell. Endocrin.* **92**, No.1, 83-90

[93] Chandler, L.A., Sosnowski, B.A., McDonald, J.R., Price, J.E., Aukerman, S.L., Baird, A., Pierce, G.E. & Houston, L.L. (1998). Targeting tumour cells via EGF receptors: Selective toxicity of an HBEGF-toxin fusion protein. *Internat. J. Cancer* **78**, No.1, 106-111

[94] Ippoliti, R., Ginobbi, P., Lendaro, E., D'Agostino, I., Ombres, D., Benedetti, P.A., Brunori, M. & Citro, G. (1998). The effect of monensin and chloroquine on the endocytosis and toxicity of chimeric toxins. *Cell. & Molec. Life Sci.* **54**, No.8, 866-875

[95] Uckun, F.M., Chelstrom, L.M., TuelAhlgren, L., Dibirdik, I., Irvin, J.D., Langlie, E.C. & Myers, D.E. (1998). TXU (Anti-CD7)-pokeweed antiviral protein as a potent inhibitor of human immunodeficiency virus. *Antimicro. Agents & Chemother.* **42**, No.2, 383-388

[96] Wang, Y., Qian, R.Q., Gu, Z.W., Jin, S.W., Zhang, L.Q., Xia, Z.X., Tian, G.Y., Ni, C.Z. (1986). Scientific evaluation of Tian Hua Fen (THF) - History, chemistry and application. *Pure Appl. Chem.* **58**, No.5, 789-798

[97] The second laboratory, Shanghai Institute of Experimental Biology. (1976). Studies on the mechanisms of abortion by trichosanthin. *Scientia Sin.* **19**, 811-827

[98] Tso, J.K., Lin, S.F. & Lin, L.Y. (1976). *Acta Zool. Sin.* **22**, 166-171

[99] Barbieri, L., Battelli, M.G. & Stirpe, F. (1993). Ribosome-inactivating proteins from plants. *Biochim. Biophys. Acta* **1154**, 237-282

[100] Endo, Y., Mitsui, K., Motizuki, M. & Tsurugi, K. (1987). The mechanism of action of ricin and related toxic lectins on eukaryotic ribosomes. *J. Biol. Chem.* **262**, 5908-5912.

[101] Endo, Y. & Tsurugi, K. (1987). RNA *N*-glycosidase activity of ricin A-chain. *J. Biol. Chem.* **262**, 8128-8130.

[102] Glück, A., Endo, Y. & Wool, I.G. (1992). Ribosomal RNA identity elements for ricin A-chain recognition and catalysis: Analysis with tetraloop mutants. *J. Mol. Biol.* **226**, 411-424.

[103] Chan, Y.-L., Endo, Y. & Wool, I.G. (1983). The sequence of the nucleotides at the α-sarcin cleavage site in rat 28S ribosomal ribonucleic acid. *J. Biol. Chem.* **258**, 12768-12770.

[104] Fernandez-Puentes, C. and Vazquez, D. (1977). Effects of some proteins that inactivate the eukaryotic ribosome. *FEBS Letts.* **78**, 143-146.

[105] Montanaro, L., Sperti, S., Mattioli, A., Testoni, G. & Stirpe, F. (1975). Inhibition by ricin of protein synthesis *in vitro*. *Biochem. J.* **146**, 127-131.

[106] Moazed, D., Robertson, J.M. & Noller, H.F. (1988). Interaction of elongation factors EF-G and EF-Tu with a conserved loop in 23S RNA. *Nature* **334**, 362-364.

[107] Hausner, T.-P., Atmadja, J. & Nierhaus, K.H. (1987). Evidence that the G2661 region of 23S rRNA is located at the ribosomal binding sites of both elongation factors. *Biochimie* **69**, 911-923.

[108] Endo, Y. & Wool, I.G. (1982). The site of action of α-sarcin on eukaryotic ribosomes. *J. Biol. Chem.* **257**, 9054-9060.

[109] Endo, Y. & Tsurugi, K. (1988). The RNA *N*-glycosidase activity of ricin A-chain. *J. Biol. Chem.* **263**, 8735-8739.

[110] Endo, Y., Glück, A. & Wool, I.G. (1991). Ribosomal RNA identity elements for ricin A-chain recognition and catalysis. *J. Mol. Biol.* **221**, 193-207.

[111] Szewczak, A.A., Moore, P.B., Chan, Y.-L. & Wool, I.G. (1993). The conformation of the sarcin/ricin loop from 28S ribosomal RNA. *Proc. Natl. Acad. Sci. USA* **90**, 9581-9585.

[112] Szewczak, A.A. & Moore, P.B. (1995). The sarcin/ricin loop, a modular RNA. *J. Mol. Biol.* **247**, 81-98.

[113] Correll, C.C., Munishkin, A., Chan, Y.-L., Ren, Z., Wool, I.G. & Steitz, T.A. (1998). Crystal structure of the ribosomal RNA domain essential for binding elongation factors. *Proc. Natl. Acad. Sci. USA.* **95**, 13436-13441.

[114] Rife, J.P., Stallings, S.C., Correll, C.C., Dallas, A., Steitz, T.A. & Moore, P.B. (1999). Comparison of the crystal and solution structures of two RNA oligonucleotides. *Biophys. J.* **76**, 65-75.

[115] Marchant, A. & Hartley, M.R. (1995). The action of pokeweed antiviral protein and ricin A-chain on mutants in the alpha-sarcin loop of *Escherichia-coli* 23S ribosomal-RNA. *J. Molec. Biol.* **254**, No.5, 848-855

[116] Barbieri, L., Gorini, P., Valbonesi, P., Castiglioni, P. & Stirpe, F. (1994). Unexpected activity of saporins. *Nature* **372**, No.6507, 624

[117] Brigotti, M., Keith, G., Pallanca, A., Carnicelli, D., Alvergna, P., Dirheimer, G., Montanaro, L. & Sperti, S. (1998). Identification of the tRNAs which up-regulate agrostin, barley RIP and PAP-S, three ribosome-inactivating proteins of plant origin. *FEBS Letts.* **431**, No.2, 259-262

[118] Nicolas, E., Beggs, J.M., Haltiwanger, B.M. & Taraschi, T.F. (1998). A new class of DNA glycosylase apurinic/apyrimidinic lyases that act on specific adenines in single-stranded DNA. *J. Biol. Chem.* **273**, 17216-17220

[119] Chen, H., Hua, L., Wang, Y., Zhao, K., Yan, M.G., Dong, Y.C. (1996). RNase activity of single chain ribosome-inactivating proteins. *Progress in Biochem. Biophys.* **23**, No.5, 453-456

[120] Mock, J.W.Y., Ng, T.B., Wong, R.N.S., Yao, Q.Z., Yeung, H.W. & Fong, W.P. (1996). Demonstration of ribonuclease activity in the plant ribosome-inactivating proteins alpha- and beta-momorcharins. *Life Sciences* **59**, No.22, 1853-1859

[121] Shih, N.R., McDonald, K.A., Jackman, A.P., Girbes, T. & Iglesias, R. (1997). Bifunctional plant defence enzymes with chitinase and ribosome inactivating activities from *Trichosanthes kirilowii* cell cultures. *Plant Science* **130**, No.2, 145-150

[122] Montfort, W., Villafranca, J.E., Monzingo, A.F., Ernst, S.R., Katzin, B., Rutenber, E., Xuong, N.H., Hamlin, R. & Robertus, J.D. (1987). The three-dimensional structure of ricin at 2.8Å. *J. Biol. Chem.* **262**, 5398-5403

[123] Katzin, B.J., Collins, E.J. & Robertus, J.D. (1991). Structure of ricin A-chain at 2.5Å. *PROTEINS: Structure, Function, and Genetics* **10**, 251-259

[124] Huang, Q., Liu, S., Tang, Y., Jin, S. & Wang, Y. (1995). Studies on crystal structures, active-centre geometry and depurinating mechanism of two ribosome-inactivating proteins. *Biochem. J.* **309**, 285-298.

[125] Monzingo, A.F. & Robertus, J.D. (1992). X-ray analysis of substrate analogs in the ricin A-chain active site. *J. Mol. Biol.* **227**, 1136-1145

[126] Xiong, J.P., Xia, Z.X. & Wang, Y. (1994). Crystal-structure of trichosanthin-NADPH complex at 1.7 angstrom resolution reveals active-site architecture. *Nature Structural Biology* **1**, 695-700

[127] Husain, J., Tickle, I.J. & Wood, S.P. (1994). Crystal-structure of momordin, a Type-I ribosome-inactivating protein from the seeds of *Momordica charantia*. *FEBS Letts*. **342**, 154-158

[128] Ferrández, A., García, J.L. & Díaz, E. (1997). Genetic characterisation and expression in heterologous hosts of the 3-(3-hydroxyphenyl)propionate catabolic pathway *of Escherichia coli* K-12. *J. Bacteriol*. **179**, No. 8, 2573-2581

[129] Dagley, S., Chapman, P.J. & Gibson, D.T. (1965). The metabolism of β-phenylpropionic acid by an *Achromobacter*. *Biochem. J.* **97**, 643-650

[130] Andreoni, V. & Bestetti, G. (1986). Comparative analysis of different *Pseudomonas* strains that degrade cinnamic acid. *Appl. Environ. Microbiol*. **52**, 930-934

[131] Coulson, C.B. & Evans, W.C. (1959). Microbiological degradation of *trans*-cinnamic acid by soil *Pseudomonas*. *Chem. Ind*. **17**, 543-544

[132] Strickland, S. & Massey, V. (1973). The purification and properties of the flavoprotein melilotate hydroxylase. *J. Biol. Chem*. **248**, 2944-2952

[133] Levy, C.C. (1967). Melilotate hydroxylase. Purification of the enzyme and the nature of the prosthetic group. *J. Biol. Chem*. **242**, 747-753

[134] Cooper, R.A. & Skinner, M.A. (1980). Catabolism of 3- and 4-hydroxyphenylacetate by the 3,4-dihydroxyphenylacetate pathway in *Escherichia coli*. *J. Bacteriol.* **143**, 302-306

[135] Burlingame, R. & Chapman, P.J. (1983). Catabolism of phenylpropionic acid and its 3-hydroxy derivative by *Escherichia coli*. *J. Bacteriol.* **155**, No. 1, 113-121

[136] Harayama, S. & Timmis, K.N. (1992) in *Metal Ions in Biological Systems* (Siegel, H. & Siegel, A., eds.) Vol. **28**, pp. 99-156, Marcel Dekker, Inc., New York

[137] Savage, D.C. (1977). Microbial ecology of the gastrointestinal tract. *Ann. Rev. Microbiol.* **31**, 107-133

[138] Van der Meer, J.R., de Vos, W.M., Harayama, S. & Zehnder, A.J.B. (1992). Molecular mechanisms of genetic adaptation to xenobiotic compounds. *Microbiol. Rev.* **56**, 677-694

[139] Lam, W.W.Y. & Bugg, T.D.H. (1997). Purification, characterisation and stereochemical analysis of a C-C hydrolase: 2-hydroxy-6-keto-nona-2,4-diene-1,9-dioic acid 5,6-hydrolase. *Biochemistry* **36**, 12242-12251

[140] Henderson, I.M.J. & Bugg, T.D.H. (1997). Pre-steady-state kinetic analysis of 2-hydroxy-6-keto-nona-2,4-diene-1,9-dioic acid 5,6-hydrolase: Kinetic evidence for enol/keto tautomerisation. *Biochemistry* **36**, 12252-12258

[141] Fleming, S.M., Robertson, T.A., Langley, G.J. & Bugg, T.D.H. (2000). Catalytic mechanism of a C-C hydrolase enzyme: evidence for a *gem*-diol intermediate, not an acyl enzyme. *Biochemistry* **39**, 1522-1531

[142] Veerapandian, B., Cooper, J.B., Sali, A., Blundell, T.L., Rosati, 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

[143] Ollis, D.L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S.M., Harel, M., Remington, S.J., Silman, I., Schrag, J., Sussman, J.L., Verschueren, K.H.G. & Goldman, A. (1992). The alpha/beta-hydrolase fold. *Protein Eng.* **5**, No. 3, 197-211

[144] Díaz, E. & Timmis, K.N. (1995). Identification of functional residues in a 2-hydroxymuconic semialdehyde hydrolase. A new member of the α/β hydrolase-fold family of enzymes which cleaves carbon-carbon bonds. *J. Biol. Chem.* **270**, No. 11, 6403-6411

[145] Kawamukai, M. Submitted (JUN-1996) to the EMBL/GenBank/DDBJ databases

[146] Ferrandez, A., Garcia, J.L. & Diaz E. Submitted (NOV-1996) to the EMBL/GenBank/DDBJ databases

[147] Blattner, F.R., Plunkett, G. III, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. & Shao Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453-1474

[148] Duncan, M., Allen, E., Araujo, R., Aparicio, A.M., Chung, E., Davis, K., Federspiel, N., Hyman, R., Kalman, S., Komp, C., Kurdi, O., Lew, H., Lin, D., Namath, A., Oefner, P., Roberts, D., Schramm, S. & Davis R.W. Submitted (NOV-1996) to the EMBL/GenBank/DDBJ databases

[149] Hofer, B., Eltis, L.D., Dowling, D.N. & Timmis, K.N. (1993). Genetic analysis of a *Pseudomonas* locus encoding a pathway for biphenyl/polychlorinated biphenyl degradation. *Gene* **130**, 47-55

[150] Nordlund, I. & Shingler, V. (1990). Nucleotide sequences of the *meta*-cleavage pathway enzymes 2- hydroxymuconic semialdehyde dehydrogenase and 2-hydroxymuconic semialdehyde hydrolase from *Pseudomonas* CF600. *Biochim. Biophys. Acta* **1049**, 227-230

[151] Menn, F.M., Zylstra, G.J. & Gibson, D.T. (1991). Location and sequence of the *todF* gene encoding 2-hydroxy-6-oxohepta-2,4-dienoate hydrolase in *Pseudomonas putida* F1. *Gene* **104**, 91-94

[152] Hessing, J.G.M. (1983). Thesis. University of Leiden, Holland.

[153] Janssen, D.B., Pries, F., van der Ploeg, J., Kazemier, B., Terpstra, P. & Witholt, B. (1989). Cloning of 1,2-dichloroethane degradation genes of *Xanthobacter autotrophicus* GJ10 and expression and sequencing of the *dhlA* gene. *J. Bacteriol.* **171**, 6791-6799

[154] Feller, G., Thiry, M. & Gerday, C. (1991). Nucleotide sequence of the lipase gene *lip3* from the antarctic psychotroph *Moraxella* TA144. *Biochim. Biophys. Acta* **1088**, 323-324

[155] Atlan, D., Gilbert, C., Blanc, B. & Portalier, R. (1994). Cloning, sequencing and characterisation of the *pepIP* gene encoding a proline iminopeptidase from *Lactobacillus delbrueckii* sub sp. *bulgaricus* CNRZ 397. *Microbiology* **140**, 527-535

[156] Rhodes, G. (1993*). Crystallography made crystal clear: A guide for users of macromolecular models*. Academic Press, Inc.

[157] Giacovazzo, C., Monaco, H.L., Viterbo, D., Scordari, F., Gilli, G., Zanotti, G. & Catti, M. (1992). *Fundamentals of Crystallography* (Giacovazzo, C., ed.) Oxford University Press, New York

[158] Drenth, J. (1994). *Principles of Protein X-ray Crystallography*. Springer-Verlag, New York, Inc.

[159] Mitchell, E., Kuhn, P. & Garman, E. (1999). Demystifying the synchrotron trip: a first time user's guide. *Structure* **7**, no. 5, R111-R121

[160] Ne'eman, Y. & Kirsh, Y. (1986). Particle accelerators – or from hunters to farmers in *The Particle Hunters*. Cambridge University Press

[161] Garman, E.F. (1996). Crystallographic data collection in *Methods in Molecular Biology - Crystallographic Methods and Protocols* (Jones, C., Mulloy, B. & Sanderson, M.R., eds.) Vol. **56**, the HUMANA Press Inc., New Jersey, U.S.A.

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

[163] Sakabe, N. (1991). X-ray diffraction data collection system for modern protein crystallography with a Weissenberg camera and an imaging plate using synchrotron radiation. *Nuclear Instruments and Methods in Physics Research, Section A* **303**, 448-463

[164] Westbrook, E.M. & Naday, I. (1997). Charge-coupled device-based area detectors. *Methods in Enzymology* **276**, 244-268

[165] Otwinowski, Z. & Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. *Methods in Enzymology* **276**, 307-326

[166] Fox, G.C. & Holmes, K.C. (1966). An alternative method of solving the layer scaling equation of Hamilton, Rollet and Sparks. *Acta Cryst.* **20**, 886-891

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

[168] QUANTA96 X-ray structure analysis user's reference, June 1996. San Diego: Molecular Simulations

[169] Driessen, H., Haneef, M.I.J., Harris, G.W. & Howlin, B. (1989). Restrain - Restrained structure-factor least-squares refinement program for macromolecular structures. *J. Appl. Cryst.* **22**, 510-516

[170] Brünger, A.T., Kuriyan, J. & Karplus, M. (1987). Crystallographic Rfactor refinement by molecular dynamics. *Science* **235**, 458-460

[171] Brünger, A.T. (1992). Free R-value - a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, No. 6359, 472-475

[172] Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., GrosseKunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J. Rice, L.M., Simonson, T. & Warren, G.L. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Cryst. Section D – Biological Crystallography* **54**, 905-921

[173] Ogata, C.M. (1998). MAD phasing grows up. *Nature Structural Biology* **5**, synchrotron supplement, 638-640

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

[175] Walsh, M.A., Evans, G., Sanishvili, R., Dementieva, I. & Joachimiak, A. (1999). MAD data collection – current trends. *Acta Cryst. Section D - Biological Crystallography* **55**, 1726-1732

[176] Steller, I., Bolotovsky, R. & Rossmann, M.G. (1997). An algorithm for automatic indexing of oscillation images using Fourier analysis. *J. Appl. Cryst.* **30**, 1036-1040

[177] Leslie, A.G.W. (1997). MOSFLM user guide for MOSFLM version 5.50. MRC laboratory of molecular biology, Cambridge

[178] Weeks, C.M. & Miller, R. (1999). The design and implementation of *SnB* v2.0. *J. Appl. Cryst.* **32**, 120-124

[179] Smith, G.D., Nagar, B., Rini, J.M., Hauptman, H.A. & Blessing, R.H. (1998). The use of SnB to determine an anomalous scattering substructure. *Acta. Cryst.* **D54**, 799-804

[180] Turner, M.A., Yuan, C.-S., Borchardt, R.T., Hershfield, M.S., Smith, G.D. & Howell, P.L. (1998). Structure determination of selenomethionyl S-adenosylhomocysteine hydrolase using data at a single wavelength. *Nature Structural Biology* **5**, 369-376

[181] Blessing, R.H. & Smith, G.D. (1999). Difference structure-factor normalisation for heavy-atom or anomalous-scattering substructure determinations. *J. Appl. Cryst.* **32**, 664-670

[182] The His-Patch Thiofusion™ Expression System. Invitrogen BV, De Schelp 12, 9351 NV Leek, The Netherlands

[183] Taylor, M.A.J., Pratt, K.A., Revell, D.F., Baker, K.C., Sumner, I.G. & Goodenough, P.W. (1992). Active papain renatured and processed from insoluble recombinant propapain, expressed in *Escherichia coli*. *Protein Engineering* **5**, No. 5, 455-459

[184] Fletcher, D.A., McMeeking, R.F. & Parkin, D. (1996). The United Kingdom Chemical Database Service. *J. Chem. Inf. Comput. Sci.* **36**, 746-749

[185] Budisa, N., Steipe, B., Demange, P., Eckerskorn, C., Kellermann, J. & Huber, R. (1995). High-level biosynthetic substitution of methionine in proteins by its analogs 2-aminohexanoic acid, selenomethionine, telluromethionine and ethionine in *Escherichia coli*. *Eur. J. Biochem.* **230**, 788-796

[186] Hoyt, J., de Fontaine, D. & Warburton, W. (1984). Determination of the anomalous scattering factors for Cu, Ni and Ti using the dispersion relation. *J. Appl. Cryst.* **17**, 344-351

[187] Verschueren, K.H.G., Franken, S.M., Rozeboom, H.J., Kalk, K.H. & Dijkstra, B.W. (1993). *J. Mol. Biol.* **232**, 856-872

[188] Morris, A.L., MacArthur, M.W., Hutchinson, E.G., & Thornton, J.M. (1992). *Proteins: Struct. Funct. Genet.* **12**, 345-364

[189] Nicolas, A., Egmond, M., Verrips, C.T., de Vlieg, J., Longhi, S., Cambillau, C. & Martinez, C. (1996). Contribution of cutinase serine 42 side chain to the stabilisation of the oxyanion transition state. *Biochemistry* **35**, 398-410

[190] Brzozowski, A.M., Derewenda, U., Derewenda, Z.S., Dodson, G.G., Lawson, D.M., Turkenburg, J.P., Bjorkling, F., Huge-Jensen, B., Patkar, S.A. & Thim, L. (1991). A model for interfacial activation in lipases from the structure of a fungal lipase-inhibitor complex. *Nature* **351**, 491-494

[191] Brok, R.G.P.M., Belandia, I.U., Dekker, N., Tommassen, J. & Verheij, H.M. (1996). *Escherichia coli* outer membrane phospholipase A: role of two serines in enzymatic activity. *Biochemistry* **35**, 7787-7793

[192] Omeir, R.L., Arreaza, G. & Deutsch, D.G. (1999). Identification of two serine residues involved in catalysis by fatty acid amide hydrolase. *Biochem. Biophys. Res. Comm.* **264**, 316-320

[193] Martinez, C., Nicolas, A., Vantilbeurgh, H., Egloff, M.P., Cudrey, C., Verger, R. & Cambillau, C. (1994). Cutinase, a lipolytic enzyme with a preformed oxyanion hole. *Biochemistry* **33**, 83-89

[194] Nandhagopal, N., Senda, T., Hatta, T., Yamada, A., Masai, E., Fukuda, M. & Mitsui, Y. (1997). Three-dimensional structure of microbial 2-hydroxyl-6-oxo-6-phenylhexa-2,4-dienoic acid (HPDA) hydrolase (BphD enzyme) from *Rhodococcus Sp.* strain RHA1, in the PCB degradation pathway. *Proc. Jpn. Acad., Ser. B* **73**, 154-157