Expression and characterisation of metalloproteins from *Mycobacterium tuberculosis*

By

Rebecca Elaine Cole

A thesis submitted in partial fulfilment of the requirements of Liverpool John Moores University for the degree of Doctor of Philosophy

This research programme was carried out in collaboration with CCLRC Daresbury Laboratory

School of Biomolecular Sciences
Liverpool John Moores University
Liverpool
L3 3AF

Molecular Biophysics Group
CCLRC Daresbury Laboratory
Warrington
WA4 4AD

February 2007

Abstract

The resurgence of tuberculosis cases world-wide over the last two decades has led to one third of the population being infected and an ever increasing number of deaths (World Health Organisation, 2006). Little is known about the pathogenicity of the infectious agent, *Tubercule bacillus*, and resistance to the key chemotherapeutic drugs is widespread. Increasing research effort aiming to curtail the spread of this disease has been aided by the work of Cole *et al.* (1998 and 2002), which provided genomic annotations of the H37Rv strain of *Mycobacterium tuberculosis*. Subsequent structural genomics projects have identified hundreds of potential targets for structure-based drug design.

The research presented in this thesis focuses on the expression and characterisation of targets from the *Mycobacterium tuberculosis* genome. Cell-free expression trials of 36 unique targets were performed. Initial screening resulted in soluble expression for 30 % of the targets and inclusion of additives, such as molecular chaperones or detergents, increased this to 67 %. Milligram quantities of protein were obtained for eleven targets. As a comparison, four targets were chosen for expression trials using an *E. coli in vivo* system. Similar results were obtained for three of the targets using the cell-free or *in vivo* expression systems. However, significant quantities of soluble Rv3545c, a cytochrome P450 125, were only produced using the *in vivo* method.

Proteins that were expressed in sufficient quantities were progressed into crystallisation trials, one of which yielded crystals suitable for X-ray diffraction. The crystal structure of Rv3628, an inorganic pyrophosphatase (Mtb-PPase), was refined to 2.7 Å resolution in space group $P3_221$. Inorganic pyrophosphatases (PPases) are ubiquitous metalloenzymes which belong to the phosphatase superfamily, and play an essential role in biosynthetic reactions (Teplyakov *et al.*, 1994). The refined crystal structure of Mtb-PPase was found to exhibit a similar overall fold and oligomeric form to existing type I PPase structures. Comparison with two recent Mtb-PPase structures, both in space group $P6_322$ (Tammenkoski et al., 2005 and Benini and Wilson, to be published), highlighted a possible pH-dependent role of His93 within the active site.

The characterisation of Rv3545c, a predicted cytochrome P450 125 (Mtb-CYP125), is also described in this thesis. Cytochrome P450s are a superfamily of haem-thiolate proteins (50 to 60 kDa) which monooxygenate hydrophobic substrates as part of electron transport chains (Nebert and Gonzalez, 1987 and Chapple, 1998). P450s have recently been implicated as novel antimycobacterial targets (Munro *et al.,* 2003).

Spectroscopy was used to confirm the cytochrome P450 annotation of Rv3545c, with the ferrous enzyme exhibiting a Soret peak at 450 nm in the presence of CO. A high-to-low spin-shift was observed by UV/visible and EPR spectroscopy, upon imidazole-inhibition of ferric Mtb-CYP125. Secondary structural elements were determined by circular dichroism (CD) to be ~ 33 % α-helix and ~ 14 % β-sheet. Finally, dark brown/red crystals of Mtb-CYP125 were obtained, but it was not possible to collect a full data set. This was primarily due to the crystals forming clusters which were impossible to separate. Despite this, weak diffraction data to 3 Å resolution were measured, and further optimisation of the crystallisation conditions may prove successful.

Contents

List of figures

Chapter 1 – Introduction

Chapter 2 – Theoretical and experimental background to protein production and characterisation

Chapter 3 – Theoretical and experimental background to protein crystallography

Chapter 4 – Production of proteins from *Mycobacterium tuberculosis*

Chapter 5 - Structure of inorganic pyrophosphatase (Rv3628) from *Mycobacterium tuberculosis*

Chapter 6 – Characterisation of cytochrome P450 125 (Rv3545c) from *Mycobacterium tuberculosis*

List of tables

Chapter 1 – Introduction

Chapter 4 – Production of proteins from *Mycobacterium tuberculosis*

Chapter 5 - Structure of inorganic pyrophosphatase (Rv3628) from *Mycobacterium tuberculosis*

Chapter 6 – Characterisation of cytochrome P450 125 (Rv3545c) from *Mycobacterium tuberculosis*

Abbreviations

ADP	Adenosine diphosphate
AMP	Adenosine monophosphate
ATP	Adenosine triphosphate
Bs-PPase	*Bacillus subtilis* inorganic pyrophosphatase
CCP4	Collaborative computer project 4
CD	Circular dichroism
CMC	Critical micelle concentration
CSA	Camphor-10-sulfonic acid
CYP	Cytochrome P450 monooxygenase
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
E-PPase	*Escherichia coli* inorganic pyrophosphatase
EPR	Electron paramagnetic resonance
ESU	Estimated standard uncertainty
HPL	*Medicago truncatula* hydroperoxide lyase (CYP74C3)
HS	High-spin haem iron
IMAC	immobilised metal ion adsorption chromatography
IPTG	isopropyl-β-D-thiogalactopyranoside
Kan	Kanamycin
kDa	Kilo Dalton
LB-Amp	Luria-Bertani media containing 100 μg/ml ampicillin
LB-Amp-Cam	Luria-Bertani media containing 100 μg/ml ampicillin and 34 μg/ml chloramphenicol
LB-Kan	Luria-Bertani media containing 35 μg/ml kanamycin
LB-Kan-Cam	Luria-Bertani media containing 35 μg/ml kanamycin and 34 μg/ml chloramphenicol
LS	Low-spin haem iron
MES	2-(N-morpholino)ethanesulphonic acid
MPD	M-phenylenediamine
MR	Molecular replacement
mRNA	Messenger ribonucleic acid

M. tb	*Mycobacterium tuberculosis*
Mtb-CYP121	*Mycobacterium tuberculosis* cytochrome P450 121
Mtb-CYP124	*Mycobacterium tuberculosis* cytochrome P450 124
Mtb-CYP125	*Mycobacterium tuberculosis* cytochrome P450 125
Mtb-PPase	*Mycobacterium tuberculosis* inorganic pyrophosphatase
MWCO	Molecular weight cut-off
NAD(P)H	Nicotinamide adenine dinucleotide (phosphate)
NCBI	National Center for Biotechnology Information
NMR	Nuclear magnetic resonance
NWSGC	North West Structural Genomics Consortium
OD	Optical density
P420	Cytochrome P420 monooxygenase
P450	Cytochrome P450 monooxygenase
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PEG	Polyethylene glycol
Pfu-PPase	*Pyrococcus furiosus* inorganic pyrophosphatase
Pho-PPase	*Pyrococcus horikoshii* inorganic pyrophosphatase
PPase	Inorganic pyrophosphatase
Pi	Inorganic phosphate
PPi	Inorganic pyrophosphate
PX	Protein crystallography
RMSD	Root mean square deviation
RNA	Ribonucleic acid
SAXS	Small-angle X-ray scattering
SDS	Sodium dodecyl sulphate
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
Sg-PPase	*Streptococcus gordonii* inorganic pyrophosphatase
Sm-PPase	*Streptococcus mutans* inorganic pyrophosphatase
S-PPase	*Sulfolobus acidocaldarius* inorganic pyrophosphatase
SR	Synchrotron radiation
SRS	Synchrotron Radiation Source
TB	Tuberculosis
TB-Amp	Terrific broth media containing 100 μg/ml ampicillin
TB-Amp-Cam	Terrific broth media containing 100 μg/ml ampicillin and

	34 µg/ml chloramphenicol
TB-Kan	Terrific broth media containing 35 µg/ml kanamycin
TB-Kan-Cam	Terrific broth media containing 35 µg/ml kanamycin and 34 µg/ml chloramphenicol
TBSGC	Tuberculosis Structural Genomics Consortium
Tev	*Tobacco etch virus*
T-PPase	*Thermus thermophilus inorganic pyrophosphatase*
tRNA	Transfer ribonucleic acid
UV	Ultra violet
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside
Y-PPase	*Saccharomyces cerevisiae* inorganic pyrophosphatase

Acknowledgements

Firstly thanks to my director of studies, Professor Samar Hasnain, for providing this opportunity and for all his guidance over the years. Also my thanks go to CCLRC, Astex Therapeutics, and Liverpool John Moores University, for providing the funding for this research.

Thanks to my second supervisor, Professor Colin Reynolds, and also to Dr John Baldwin, both from JMU, for all their advice and positive words.

To my (unofficial) third supervisor, who I 'adopted' during my second year, Dr Toni Georgiou. My time in the SBL would have been much harder without her many words of wisdom (especially about gloves), particularly during my moments of panic. Also, thanks to her for reading several of my thesis chapters…I'm not sure what she'll do for entertainment in the pub now!

To everyone in the Molecular Biophysics Group, Drs Mike Hough, Svetlana Antonyuk, Günter Grossmann, Richard Strange, Mark Ellis, and Mark Barrett, for their help over the years. In particular to Sveta, for her crucial help during crystallography and for her refusal to give in during data collection, even in the face of multiple salt crystals. Also many thanks to Mike for reading my thesis, and everything else that I've written during this PhD. I still owe you that pint…

To Drs Pamela Williams and Hayley Angrove from Astex Therapeutics for answering my many P450 questions and for reading several of my chapters. To Dr Radek Kowalczyk at the EPSRC National Centre for EPR for his help during EPR data collection and to Mr Alan Brown for his guidance during CD measurements. Thanks also to Drs Richard Hughes and Eric Belfield from the John Innes Centre, for their enthusiasm and kindness, and for providing vast quantities of protein during my work with plant P450s.

I am indebted to Mr Takayoshi Matsuda, for his substantial assistance throughout my visits to RIKEN and for making me feel welcome in Japan. Thanks to Professors Shigeyuki Yokoyama and Takanori Kigawa, Ms Eiko Seki, Dr Takashi Yabuki, and

many others at the Protein Research Group of the RIKEN Yokohama Institute, for their help and kindness throughout. Also thanks to Ms Natsuko Matsuda, Mrs Akiko Kikuchi, and Ms Yukako Miyata, for making my experience in Japan truly memorable.

To Mrs Theresa Hillon and Mrs Margaret Jones for always being on hand to help with emergencies in the lab, and for the many chats. Also, to Drs Miroslav Papiz and Katie Evans for sharing their molecular biology knowledge with me. To the many people who I've met during my PhD, many of whom I know only by face, for making Daresbury a happier place to work. In particular to Mrs Glenys McBain, for her words of reason and for always being cheerful, to Mr Jimmy McCabe, and Drs John Warren and Neil Telling.

To my family and friends, who have kept me sane during the dark days and have also been there to share the many good times. To the Mediterranean complement of team MBG, Drs Kostas Paraskevopoulos, Tina Geraki, Michele Cianci, and (soon-to-be) Dr Sonia Barbieri, all of whom have become good friends and allies over the years. It's been an absolute pleasure to work with you guys and our long chats, especially at 'Daresbury beach', will be sadly missed.

Having thought they could finally stop funding me, I am still much indebted to my parents for their financial support over the last few months...I promise I'll get a real job now!! More importantly though, for the unconditional emotional support which they, and my sister Laura, have always provided. Also, thanks to Mia and Jaxx.

And finally, to my partner Matt Edwards. Having spent many a sleepless night contemplating what to say at this point, I am now lost for words. I can't emphasise how important his support has been over the last few years, so I'll just say thanks for everything. For listening to me talk for hours about work, for always knowing what to say, for making me realise what's important, and for bailing me out during the last few months of my write up. Thank you.

Declaration

All work presented in this thesis is the original work of the author, except where acknowledged by reference. No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this university or any other institution of learning.

Signed ....

Dated ....

Preface

This thesis is a report of original research undertaken by the author and is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy to Liverpool John Moores University.

The expression and characterisation of metalloproteins from the pathogen *Mycobacterium tuberculosis* are the focus of research presented in this thesis. The expression of 36 targets was attempted using a bacterial cell-free protein expression system. As a comparison, the expression of four targets was also attempted using an *E. coli in vivo* system. The crystal structure of Rv3628, an inorganic pyrophosphatase, which was successfully expressed using the cell-free system, was refined to a resolution of 2.7 Å and is described in this thesis. Finally, a cytochrome P450 125 (Rv3545c), expressed using the *in vivo* system, was characterised by bioinformatics, UV/visible spectroscopy, circular dichroism, and electron paramagnetic resonance. Crystallisation trials were also undertaken, yielding weakly-diffracting multiple crystals.

The research presented in this thesis was performed predominantly at CCLRC Daresbury Laboratory . All cell-free protein expression studies were carried out at the Protein Research Group of the RIKEN Yokohama Institute, Japan. Electron paramagnetic resonance experiments were conducted at the EPSRC National Centre for EPR Spectroscopy, Manchester.

Structure of the thesis:

Chapter 1: A brief introduction to the disease tuberculosis and the causative agent, *Mycobacterium tuberculosis*, together with background information regarding inorganic pyrophosphatases and cytochrome P450s

Chapter 2: A theoretical and experimental background to the production of proteins and their subsequent characterisation used throughout this thesis

Chapter 3: A review of the theoretical and experimental background regarding protein crystallography

Chapter 4: Describes the expression trials of protein targets from *Mycobacterium tuberculosis* using both *E. coli*-based cell-free and *in vivo* systems and provides a brief comparison between the two systems

Chapter 5: Describes the crystallisation and structure determination of an inorganic pyrophosphatase (Rv3628) from *Mycobacterium tuberculosis* to 2.7 Å resolution

Chapter 6: Describes the characterisation of cytochrome P450 125 (Rv3545c) from *Mycobacterium tuberculosis* by bioinformatics, UV/visible spectroscopy, circular dichroism, and electron paramagnetic resonance. This chapter also details crystallisation of this protein and the subsequent data collection from weakly-diffracting multiple crystals

Chapter 7: Final conclusions regarding the entire thesis and suggestions for future work

Chapter 1 – Introduction

1.1 *Mycobacterium tuberculosis*

1.1.1 Tuberculosis

Tuberculosis is a chronic infectious disease caused by the pathogen *Tubercle bacillus*, which infects one in three people and accounts for more than two million deaths each year. Due to the highly infectious, airborne nature of the disease, the World Health Organisation (WHO) predicts that by 2020 nearly a billion people will be infected, leading to two hundred million cases and thirty-five million deaths worldwide. In 1993, as deaths from this single pathogen totalled more than those from malaria, diarrhoea, HIV/AIDS, and tropical diseases combined (Evans, 1998), the WHO declared a state of global emergency.

Despite these devastating statistics, very little is known about the pathogenicity of this mycobacterium. Until the late 1990s tuberculosis was widely regarded as a disease of the poor (Evans, 1998). However, the sharp incline in HIV and AIDS cases in the Western world has accelerated the spread of this disease (WHO, 2000). Although a large proportion of the world's population is infected with TB, only 5 - 10 % become ill, due to the host's immune system attacking the mycobacterium and forcing it to remain dormant. However, in immuno-suppressed hosts, rapid proliferation of the disease occurs and is the leading cause of death in HIV patients.

The Bacille Calmette-Guerin (BCG) vaccine has prevented the number of deaths from tuberculosis rising further, conferring protection by inoculating the patient with a live attenuated strain of *Mycobacterium bovis*. This closely related species provides immunological protection from *Mycobacterium tuberculosis* in the same way that inoculation with cowpox prevents the emergence of smallpox (Evans, 1998).

Until recently, development of anti-mycobacterial compounds has been fairly stagnant, medication instead relying upon five key chemotherapeutic drugs: isoniazid, rifampicin, pyrazinamide, ethambutol/streptomycin, and capriomycin. The use of these drugs as a first line defence against tuberculosis has led to widespread resistance, due in part to incorrect drug regimes. Despite the emergence of resistant tuberculosis strains, the Direct Observation of Treatment (DOTs) scheme has a success rate of > 95 % and costs just US$ 11 per 6 month supply in developing countries. The scheme encompasses not only drug therapy, but also encourages political commitment, provides microbiology detection services, and monitors drug administration to prevent further resistance occurring. By 2015, the WHO aims to have reduced cases by 50 % compared with 1990 levels, with complete eradication by 2050 (WHO, 2006).

Over the last few years, tuberculosis research has come under the spotlight with the completion of the *M. tb* genome sequence, strain H37Rv (Cole *et al.*, 1998). The information gained from the genome, together with subsequent structural genomics projects, will improve our biological understanding of this pathogen and allow us to design tailored anti-mycobacterial drugs.

1.1.2 The *Mycobacterium tuberculosis* genome

Our knowledge of *Mycobacterium tuberculosis* has been significantly advanced by the work at the Wellcome Trust Sanger Institute in Cambridge, where the research group headed by S. T. Cole, generated a significant amount of data regarding the pathogens genome. This was achieved by a combination of systematic sequence analysis from cosmids and BACS for large insert clones, together with random small-insert clones from a whole genome shotgun library. They identified large areas of repetition within the genome, together with many insertion sequences, more than fifty of which were intergenic or non-coding and localised near to tRNA genes. They postulated that this prevents key genes from being inactivated (Cole *et al.*, 1998).

The complete genome sequence, published in 1998, was that of the most characterised tuberculosis strain, H37Rv (Cole *et al.*, 1998). The genome, unusually rich in guanidine and cytosine bases (65.6 % compared with 50.8 % for *E. coli*), encompasses 4, 411, 529 base pairs and 3, 974 genes, the second largest bacterial genome sequence currently available. 3, 924 genes encode for peptides and 50 for stable RNA. The low abundance of adenine & thymine residues is reflected in the organism's codon bias, highlighted by the abnormally high frequency with which GTG start codons are observed (35 %

compared with only 9 % in *Bacillus subtilis*). Greenacre's correspondence analysis showed a preference for G/C-rich amino acids such as alanine, glycine, proline, arginine, and tryptophan, and this GC content has been shown to be fairly constant throughout the genome (Cole *et al.,* 1998).

This phenomenon of codon bias occurs due to there being sixty-four different codons for only twenty amino acids. Therefore there is no requirement for all of the codons, all of the time. The third position base is usually insignificant and can be interchanged forming synonymous residues. Such residues are not used with the same frequency due to translational selection and mutational bias.

Functional analysis of the tuberculosis genome allowed Cole's group to categorise many of the genes with regards to their assumed role. The eleven categories for which they identified were 1 - Virulence, detoxification, and adaption (91 genes); 2 - Lipid metabolism (225 genes); 3 – Information pathways (207 genes); 4 - Cell-wall and cell processes (516 genes); 5 - Stable RNAs (50 genes); 6 - Insertion sequences and phages (137 genes); 7 - PE and PPE proteins (proteins with N-terminal Pro-Glu and Pro-Pro-Glu motifs, respectively) (167 genes); 8 - Intermediate metabolism and respiration (877 genes); 9 - Proteins of unknown functions (606 genes); 10 - Regulatory proteins (188 genes); and 11 - Conserved hypothetical proteins (910 genes).

This research has shown that a significantly large proportion of coding capacity from this mycobacterium is taken up by the biosynthesis of lipogenesis and lipolysis enzymes, including an extensive range of Cytochrome P450s. Two new families of glycine-rich proteins, PE and PPE, containing repetitive sequences which may represent antigenic variation, were also encoded for. An unpredicted observation of their work was a potential mechanism which allows *Mycobacterium tuberculosis* to present the immune system with a "moving target", by altering the expression pattern of a number of short peptides, thus evading destruction (Cole *et al.,* 1998).

Re-annotations of the H37Rv genome (now 4, 411, 532 nucleotides) published in 2002, identified twenty-two new protein-coding sequences with predicted functions, together with a further sixty of unknown function (Cole *et al.,* 2002). No new RNA genes were identified (**table 1.1**). The bioinformatics tools, BLASTP and FASTA (Pearson and

Lipman, 1988) were exploited to manually re-analyse the protein-coding sequences within the original sequence.

During this re-analysis, the group detected new sequences containing the appropriate GC content, correlation scores, and codon usage found in coding sequences. Further inspection employing the codon-usage program AMIGA (Automatic Microbial Genome Annotation), identified potential frameshifts and coding sequences. See **tables 1.1 to 1.2** for a description of *M. tb* genomic annotations.

Table 1.1: Annotations of the *Mycobacterium tuberculosis* genome. Genes grouped by their proposed function (Cole *et al.*, 1998 and 2002). [1]Cole *et al.*, 1998. [2]Cole *et al.*, 2002.

Table 1.2: Changes to the functional classification of *Mycobacterium tuberculosis* genes following re-annotation by Cole *et al.* in 2002.

1.1.3 Identification of essential genes

1.1.3a Virulence, growth, and survival genes

The identification of genes essential for *M. tb* virulence, growth, or survival, during infection will undoubtedly aid rational drug design in the fight against tuberculosis. A key paper in the determination of genes essential for mycobacterial survival was published in 2003 by Sassetti and Rubin (Sassetti and Rubin, 2003). This group created a library of mutants from the *M. tb* (H37Rv) genome, which were grown on agar plates for several weeks (*in vitro* pools) and also used to infect mice (*in vivo* pools) for the same period. Surviving bacteria were collected from the mice and re-plated onto agar, before being compared with the *in vitro* pools using transposon site hybridisation (TRASH). TRASH enabled the identification of all mutated genes present within the *M.*

tb cells prior to, and after infection. Those which were not present in the *in vivo* pools were deemed to be essential for *M. tb* survival. The results of this study are summarised in **table 1.3**, together with those from previous work describing genes essential for *in vitro* survival (Sassetti *et al.*, 2001).

1.1.3b Drug resistance genes

Understanding the means by which *M. tb* confers resistance to chemotherapeutics, will enable the development of novel treatments. Resistance occurs frequently in tuberculosis due to the compartmentalisation of infection within the host, which prevents multiple-chemotherapeutics from functioning simultaneously, thereby resulting in monotherapy (Gillespie, 2002). Dessen *et al.* (2005) described the three mechanisms whereby such resistance can occur: inactivation of the antibiotic by modification; mutagenesis of key residues, resulting in modification of the macromolecular target; and promotion of antibiotic efflux from the cell. The most commonly used antibacterials target protein synthesis, nucleic acid replication and repair, and cell wall biosynthesis mechanisms (Dressen *et al.*, 2005).

Resistance to the antimycobacterial agent, streptomycin, was found to be due to mutations within the 16S RNA or the S12 protein, both involved in protein synthesis (Dressen *et al.*, 2005). The identification of these proteins as drug targets was determined by the high resolution structure of the 30S ribosomal subunit, in complex with streptomycin (Carter *et al.*, 2000).

Missense mutations within a conserved region of the rpoB gene, which encodes the β-subunit of RNA polymerase, account for 97 % of all rifampicin-resistant *M. tb* isolates, as determined by RFLP analysis (Yuen *et al.*, 1999). Another mechanism by which *M. tb* acquires resistance is via the inactivation of an enzyme required by the drug for activity. The drug isoniazid requires catalase for activation and mutations within this gene, KatG, confer resistance (Gillespie, 2002). Finally, pyrazinamidase is the target for the chemotherapeutic, pyrazinamide. Resistance to this drug is due to mutations within the enzyme which prevent conversion of pyrazinamide to its active form, pyrazinoic acid (Gillespie, 2002).

Table 1.3: Genes found to be essential for *Mycobacterium tuberculosis* survival *in vivo*^{1} (Sassetti and Rubin, 2003) and *in vitro*^{2} (Sassetti *et al.*, 2001).

1.1.4 Structural biology consortia

Several research projects dedicated to elucidating the complex structure-function relationships of the pathogen *Mycobacterium tuberculosis* are spread throughout the world. Consortia have been formed on varying scales to encourage the flow of information and accelerate the discovery of new drugs. Such groups within North America include the Tuberculosis Trials Consortium (TBTC); the Tuberculosis Results Action Consortium (TRAC); and the Tuberculosis Epidemiological Studies Consortium (TBESC). Closer to home are the XMTB German *Mycobacterium tuberculosis*

Structural Genomics project; the French Pasteur Institute; and the North West Structural Genomics Consortium (NWSGC), of which Daresbury Laboratory is a key member, together with the Universities of Manchester, Leeds, Liverpool, Liverpool John Moores, Astra Zeneca, and Astex Therapeutics, Cambridge. This consortium alone has over forty targets for structural characterisation.

International efforts include the Global Alliance for Tuberculosis which operates in New York, Brussels, and Cape Town and, the Tuberculosis Structural Genomics Consortium (TBSGC). The latter comprises more than 230 members from 31 organisations, covering 13 countries, including the UK. To date, fifty crystal structures have been posted on the consortium website. However, with the overall goal of characterising 400 potential drug targets, work is far from finished (Goulding *et al.,* 2002). See **table 1.4** for list of recent crystal structures available in the Protein Data Bank.

PDB Code	Predicted function	PDB Code	Predicted function
2HH7	CSOR	2AQ8	Enoyl-ACP (coA) reductase
2HHI	MPT64	2A6P	Hypothetical protein (Rv3214)
2NYX	Hypothetical protein (Rv1404)	2CDN	Adenylate kinase
2NQT	Hypothetical protein (Rv1652)	1ZEL	Hypothetical protein (Rv2827c)
2NTN	MabA	2CIG	Dihydroflorate reductase
2NV6	InhA	2GDN	β-lactamase
2EV1	1264N	2G2D	pdyO-type ATP cobalamin adenosyltransferase
1IJ5	CYP121	2BYO	Lipoprotein LPPX Rv2945c
2IXC	C3’ C5’ carbohydrate epimersae rmcc	2G38	PE/PPE complex
2ITW	Shikimate kinase	2BIP	EPSP synthase
2844	Hypothetical protein (Rv2844)	1V5P	LigD ligation domain
2GKM	TrHbn	2FHG	Proteasome
2C21	Hypothetical protein (Rv0130)	2A75	Acyl-coA carboxylase ACCDS
2DTF	Threonine synthase	1XXX	Dihydrodipicolinate synthase dapA Rv2753c

Table 1.4: Crystal structures from *Mycobacterium tuberculosis* published during 2006 (http://www.rcsb.org).

A paper by Terwilliger *et al.* in 2003 highlighted a few of the recent successes, ranging from a glyoxylate pathway protein, essential for bacterial survival within the macrophage (Sacchettini *et al.*, 2000 and Smith *et al.*, 2003); to a P450 which provides a good model for novel anti-TB drugs (Mowat *et al.*, 2002); and an iron-metabolic enzyme involved in NAD biosynthesis (Bossi *et al.*, 2003).

Clare Smith and James Sacchettini's research group from the Department of Biochemistry and Biophysics, Texas A & M University, successfully characterised two key enzymes in the glyoxylate shunt pathway. These proteins, isocitrate lyase, Rv0467-icl (Sacchettini *et al.*, 2000), and malate synthase, Rv1837c-glcB (Smith *et al.*, 2003), become upregulated when bacteria shift to preferentially utilising substrates generated by β-oxidation of fatty acids (Honer *et al.*, 1999) and during macrophage infection (Graham *et al.*, 1999). Hence, these enzymes are essential for bacterial survival and persistence within the activated macrophage (McKinney *et al.*, 2000) and are extremely attractive targets for new anti-mycobacterials.

As previously described, the *Mycobacterium tuberculosis* genome encodes a wide array of lipid-metabolising Cytochrome P450s, which oxidise fatty acids, sterols, and steroids (Porter and Coon, 1991). Generally prokaryotes encode for only a few, if any, P450s which led to Mowat *et al.* from the University of Edinburgh postulating that the number of P450s corresponds to the importance of lipid metabolism within the pathogen (Mowat *et al.*, 2002). Subsequently they crystallised the mycobacterial CYP121 (gene Rv2276), which binds tightly to azole-based antifungals such as miconazole and clotrimazole. Such drugs potently inhibit P450s in the nanomolar range, and are hence obvious candidates for anti-mycobacterial drug development (Mowat *et al.*, 2002). It was not possible to solve the structure using molecular replacement, despite obtaining diffraction data to 1.06 Å, which represents a common problem associated with P450s as they tend to lack sequence homology between species. Instead, Multiple Isomorphous Replacement with Anomalous Scattering (MIRAS) was successfully used (Leys *et al.*, 2002).

An Italian research group from the University of Pavia, working on proteins thought to be associated with NAD biosynthesis and iron metabolism, have recently determined the crystal structure of FprA (gene Rv3106) in both its oxidised (at 1.05 Å reoslution) and reduced (at 1.25 Å resolution) forms in complex with NADP$^+$. FprA has been classified as a mycobacterial oxidoreductase due to its significant sequence identity with mammalian and yeast adrenodoxin reductase. This highly structurally conserved family of enzymes plays a role in either iron metabolism or in Cytochrome P450 reductase activity, catalysing the transfer of reducing equivalents from NADPH to a protein acceptor (Bossi *et al.*, 2002).

1.1.5 Other research efforts

Following on from Cole *et al's* genome sequence which highlighted the large percentage (~ 18 %) of coding capacity dedicated to proteins involved within the cell wall, much of today's research revolves around such processes. It is well established that the thick and waxy, hydrophobic cell envelope is formed of four types of polymers: peptidoglycan, arabinogalactan, mycolic acids, and lipoarabinomannan (Evans *et al.,* 1998). This forms a permeability barrier around the mycobacterium protecting it from a number of antibiotic drugs (Brennan, 1995). By targeting proteins which the pathogen relies upon for protection, such as those involved in cell envelope synthesis or drug modifying/efflux enzymes, it may be possible to suppress the pathogenesis of this organism.

A preliminary report of progress from the Centre of Proteomics and Genomics at the University of California, Los Angeles, outlined their tuberculosis protein targets and current status. The five classes for which they are interested are: extracellular proteins potentially involved in *M. tb* pathogenicity; iron-regulatory proteins essential for pathophysiology; functionally related proteins of known anti-TB drugs; mycobacterium-specific proteins likely to be involved in virulence and pathogenicity; and proteins containing predicted novel folds (Goulding *et al.,* 2002).

A group at UCLA have obtained a crystal structure (to 2.4 Å) of one of their targets, Rv2220, which encodes for glutamine synthase (Gil, *et al.,* 1999). Previous research by Harth (1999) and Tullius (2001) has shown this enzyme to be involved in the early stages of infection and they conclude that this may play a role in the synthesis of unique pathogenic cell-wall polymers.

Another success has been achieved by a different group at UCLA, at the UCLA-DOE Laboratory of Structural Biology and Molecular Medicine (Anderson *et al.,* 2001). The group solved the structure of a major secretary protein, mycolyl transferase antigen 85B, which together with Antigen 85A and 85C (Sacchettini *et al.,* 2000) catalyses the transfer of mycolic acids within the pathogen. These form major components of the mycobacterium's cell wall and are unique to mycobacteria, thus representing attractive drug targets. Experiments by Horwitz *et al.* in 1995 demonstrated that vaccination of guinea pigs with purified *M. tb* Antigen 85B induces considerable immunological protection against *M. tb* aerosol bacterium. Furthermore, inoculation with a recombinant form of the existing *M. bovis* BCG vaccine, expressing the *M. tb* A85B protein, induces a stronger immunity than the existing vaccine (Horwitz, *et al.,* 2000).

1.2 Inorganic pyrophosphatases

1.2.1 Phosphatases

Phosphatases catalyse the removal of phosphate groups, attached to proteins by the action of kinases. Such cycles of phosphorylation and dephosphorylation provide a reversible regulation of many metabolic pathways, according to cellular requirement (Hames and Hooper, 2000 and Busam *et al.*, 2006). Regulation of pathways such as glycogen synthesis, occurs by reversible covalent modification of the enzyme by attachment of a phosphate molecule to a hydroxyl group, often in an ATP-dependent manner. Such phosphoryl transfer alters the enzyme's tertiary structure, resulting in either up or down regulation of activity (Hames and Hooper, 2000). This is reversed by the action of phosphatases, which cleave the bond between the phosphate and enzyme, thereby releasing free phosphate. Some kinases and phosphatases act upon specific residues within the enzyme, such as threonine, serine, tyrosine, and histidine.

This ubiquitous family can be grouped dependent upon substrate specificities, catalytic mechanisms, and amino acid sequences (Dombradi, 2002). The four groups are: phosphoprotein phosphatases; metal-ion-dependent proteins; tyrosine protein phosphatases; and histidine protein phosphatases. The most characterised of all phosphatases, tyrosine phosphatase, regulates the signal transduction pathways involving tyrosine phosphorylation. These enzymes have been implicated in the development of cancer, diabetes, rheumatoid arthritis and hypertension (Van Montfort *et al.*, 2003). A less well known member of this family is histidine acid phosphatase, which functions optimally at low pH, to hydrolyse phosphate esters (Van Etten *et al.*, 1991). Two conserved catalytically important histidines, are thought to form a phosphohistidine intermediate, and to act as a proton donor (INTERPRO).

1.2.2 Inorganic pyrophosphatases

Inorganic pyrophosphatase, PPase, (EC.3.6.1.1) belongs to the phosphatase family of enzymes and catalyses the hydrolysis of the high energy compound, pyrophosphate (PPi), to orthophosphate (Pi), see **equation 1.1** (Butler, 1971 and Matthews *et al.*, 2000):

Equation 1.1:

$$P_2O_7^{-4} + H_2O \longrightarrow 2HPO_4^{-2}$$

PPi formation occurs when ATP is hydrolysed to adenosine monophosphate (AMP) during many ATP-dependent biosynthetic reactions (Voet *et al.*, 1999). In addition to hydrolysing PPi phosphonanhydride bonds (Chen *et al.*, 1990), PPase also mediates oxygen exchange between inorganic phosphate and water (Lahti, 1983). These enzymes function as part of many biosynthetic reactions, such as protein/DNA/RNA synthesis (Kankare *et al.*, 1996) and tRNA charging (Liu *et al.*, 2004), and may also be involved in the copying of DNA molecules during chromosome duplication (Lahti, 1983 and Salminen *et al.*, 1995).

PPases are ubiquitous enzymes, having been identified in virtually every organism (Teplyakov *et al.*, 1994). They are soluble proteins, predominantly found in the cytosol. Evolutionary analysis of PPase sequences from different organisms identified two distinct families: type I, which include most known PPases; and the less common type II family, which are found in Bacillus subtilus, *Methanococcus jannaschii*, and several *Streptococcus* strains (Cooperman *et al.*, 1992, Young *et al.*, 1998 and Sivula *et al.*, 1999). The two families are evolutionarily distinct and hence share no sequence homology (Tammenkoski *et al.*, 2005).

Type I PPases are well characterised and can be further divided into two groups with the prokaryotes forming hexamers of approximately 120 kDa, and the eukaryotes existing as 60 to 70 kDa homodimers. Identity between prokaryotic PPases is reasonably high (~ 45 %), however between the two groups is below 25 % (Teplyakov *et al.*, 1994). A further group of membrane-bound PPases exist in plants and some bacteria, which function as reversible proton pumps, whilst hydrolysing and synthesising PPi, and share no sequence homology with the two families previously mentioned (Sivula *et al.*, 1999).

1.2.3 Importance of inorganic pyrophosphatases

Nucleotide triphosphate-dependent biosynthetic reactions such as nucleic acid polymerisation, coenzyme synthesis, and amino acid activation, result in elevated levels of cellular PPi. High levels of PPi can result in cellular toxicity, thus PPase plays an essential role in controlling these potentially toxic levels (Butler, 1971). The cytosolic PPase-dependent reaction shifts the equilibration constant towards biosynthesis (Kornberg, 1962) and has been shown to be essential for *E. coli* (Chen *et al.*, 1990) and S. cerevisiae (Lundin *et al.*, 1991) viability.

An increase in expression of PPase from *Legionella pneumophila* was found to occur in response to environmental stimuli during intracellular infection of macrophage-like cells (Kwaik, 1998). Further experiments by Triccas and Gicquel (2001) attempted to discern whether such a connection existed during infection with *Mycobacterium tuberculosis*, however they could find no such link. This may be due to the differences in intracellular environments which are encountered by the two pathogens (Kwak *et al.*, 1999 and Triccas and Gicquel, 2001)

Despite this, a recent review of the transcriptional response of *Mycobacterium tuberculosis* to different drugs and growth-inhibitory conditions identified an up-regulation of the PPase gene (Boshoff *et al.*, 2004). Using microarray profiling, the study identified clusters of genes which were co-ordinately regulated under various stress conditions. The PPase gene belongs to the gene cluster (GC-71), implicated in ribosomal architecture and translation, and was found to be induced in response to inhibition of translation. This suggests an important role of this enzyme, however a further study by Sassetti and Rubin (2003) did not identify an essential requirement for *Mycobacterium tuberculosis* PPase during *in vivo* infection in mice. This alone however does not exclude the possibility of an important role for PPase in *Mycobacterium tuberculosis*.

1.2.4 The role of metals in pyrophosphatase activity

Cooperman and Chiu (1973) first identified the essential role of divalent metal cations in PPase catalysis, however the mechanism by which these metals exert activity has only recently been outlined. Type I PPases exhibit a preference for Mg^{2+}, decreasing in efficiency with Zn^{2+}, Co^{2+}, Mn^{2+}, and Cd^{2+} at different pH's (Lahti and Kolakowski *et al.*, 1990), which bind with a micromolar affinity (Fabrichniy *et al.*, 2004). Type II enzymes however, prefer manganese or cobalt ions which bind with a greater, nanomolar affinity (Merckel *et al.*, 2001 and Fabrichniy *et al.*, 2004), with zinc acting as both a partial activator and inhibitor (Zyryanov *et al.*, 2004). Suggestions for these differences are discussed further in section 1.2.8. Despite this preference for magnesium, manganese binds more tightly to the PPase (type I) active site, subsequently intensifying substrate/product binding affinity (Cooperman, 1981). In the absence of metal cations, only 5 % of the catalytic sites are predicted to be occupied by PPi (Janson *et al.*, 1979). PPases are also relatively thermostable, particularly when cations are bound (Ichiba *et al.*, 1998).

Significant contributions to the role of metals in PPase catalysis were conducted in Yeast and *E. coli* (Rapoport, 1973, Baykov, 1974, Cooperman, 1981, and Knight, 1984). This demonstrated E-PPase's dependence upon four metal ligands for activity, whilst the larger Y-PPase utilizes only three. Metal ions (such as magnesium) activate the enzyme and neutralise substrate net charge, thus forming part of the active substrate MgPPi, and also stabilise the transition state (Cooperman, 1982 and Baykov, 1996, Samygina, 2001). The rate of catalysis was found to be proportional to the concentration of MgPPi and free Mg^{2+} (Moe and Butler, 1972 and Rapoport *et al.*, 1972).

1.2.5 Mechanism

The mechanism by which PPase functions differs between type I and II enzymes (Fabrichniy *et al.*, 2004), as demonstrated by the varying affinity for metal cofactors described earlier. Little is known about the type II mechanism and so the type I PPase mechanism is discussed further here.

Initial understanding of the PPase catalytic mechanism resulted from work by Cooperman *et al.* (1982), which predicted that a general base activation of an attacking nucleophilic water molecule within the active site, together with the activation of the phosphoryl leaving group through the formation of a metal ion complex, led to a general acid catalysis. Further work by Baykov in 1992 identified two potential catalytic pathways, with magnesium acting as a cofactor (Baykov and Shestakov, 1992). Two activating Mg^{2+} ions were found to bind primarily to the active site, followed by PPi, and a third substrate metal ion. The fourth Mg^{2+} was found to bind loosely.

More recently, a detailed structure-based model for PPase mechanism has been described in Yeast by Heikinheimo *et al.* (1996), the resulting mechanistic model is described here and shown in **figure 1.1**. This was produced by docking a transition state model into the PPase-product structure (PDB ID: 1WGJ). As described before, hydrolysis of the P_2O_7-Mn_2 substrate occurs through attack by a potent nucleophile. Through site-directed mutagenesis and extensive biochemical characterisation, they identified the most plausible general base candidate for nucleophilic attack of PPi, to be a hydroxide ion (water 1, **figure 1.1**). This is further substantiated by the structure of fluoride-inhibited Y-PPase which also positions the hydroxide ion in close contact with the phosphoryl group, P2 (Heikinheimo *et al.*, 2001). Stabilisation of this ion occurs through interaction of its lone pairs with two active site metal ions (Mn1 and Mn2, **figure 1.1**), together with a hydrogen bond between the hydroxide's hydrogen and a side-chain oxygen from Asp117 (Y-PPase numbering). They proposed that the negatively charged hydroxide attacks the electrophilic phosphorous group, P2, leading to its dissociation from the Mn3 metal. The suggestion that P2 dissociates before P1 was made due to its lower binding affinity, together with its close contacts with the proposed nucleophile (water 1, **figure 1.1**). This however, contradicts previous work by Harutyunyan *et al.* (1996).

Finally, during the P-O-P hydrolysis step, they postulate that a water (water 6, **figure 1.1**) coordinated to Mn3, acts as a general acid, donating a proton to an oxygen on the leaving phosphorous group (P1, **figure 1.1**). The potential for the Mn3-coordinated residues, Arg78, Lys193, and Tyr192, to act as general acids was disregarded due to retention of activity in E-PPase containing mutations in homologous residues (Lahti and Pohjanoksa *et al.*, 1990).

Figure 1.1: Schematic representation of the PPase (type I) mechanism of catalysis, as proposed by Heikinheimo *et al.* (1996) through computational docking of Y-PPase. Lines represent hydrogen bonding (grey) and metal coordination (dashed), and arrows indicate the proposed flow of electrons. Key residues are highlighted, as are the leaving (P1) and electrophilic (P2) phosphoryl groups of pyrophosphate, and the metal groups (Mn1-4). See text for a description of the mechanism.

1.2.6 Inhibition

1.2.6a Calcium: A natural inhibitor

The role of calcium in cellular regulation is well established (Samygina *et al.*, 2001) and is known to act as a natural inhibitor of PPases, competing with the activating metals for position within the active site, as demonstrated in Y-PPase (Ridlington and Butler, 1972 and Butler and Sperow *et al.*, 1977) and E-PPase (Avaeva *et al.*, 1998 and Samygina *et al.*, 2001). Two calcium ions were found to bind to E-PPase in the absence of substrate (at positions Mn1 and Mn2, **figure 1.1**), one of which could be replaced by magnesium, demonstrating the different binding affinities of these two sites (Avaeva *et al.*, 1998). The higher affinity calcium binding site was found to coincide with the lower affinity magnesium/manganese site (Mn2, **figure 1.1**). In the presence of pyrophosphate and calcium, no activity was observed even at excessive concentrations of the ion (10 mM), implying that calcium cannot function as a PPase activating metal.

Subsequent high resolution structures of E-PPase in complex with Ca^{2+} and CaPPi (Samygina *et al.*, 2001) located three calcium atoms within the active sites (Mn1 - Mn3, **figure 1.1**), coordinated to the same residues as described for magnesium/manganese (section 1.2.5). All calcium ions were found to bind more strongly in the presence of PPi, with the third calcium binding with very weak affinity in the absence of substrate. Calcium and magnesium/manganese were also found to exhibit the same coordination with PPi, as described in section 1.25. Calcium inhibition was found to occur due to the unhydrolysable CaPPi competing with MgPPi for position within the active site, and also due to the inability of Ca^{2+} at site Mn2 to activate an attacking nucleophile (water 1, **figure 1.1**), required for PPi hydrolysis (section 1.2.5).

1.2.6b Other inhibitors

A less common cellular inhibitor, fluoride, was found to reversibly inactivate PPases at millimolar concentrations (Smith, 1970 and Pinkse *et al.*, 1999). Fluoride was found to rapidly bind E-PPase in the presence of MgPPi, followed by a slow decline in the inhibition rate (Baykov *et al.*, 2000). Also, fluoride is known to bind with higher affinity in the presence of substrate. In the presence of 1 mM sodium fluoride, E-PPase activity was reduced to 9 %, with less then 0.01 % activity when the concentration was increased to 10 mM (Josse, 1966). Extremely high concentrations of guanidine-HCl was also found to inhibit E-PPase, with activity reduced to 36 % and 3 % at a concentration of 1 M and 2 M, respectively (Josse, 1966).

1.2.7 Structure

As of November 2006, 36 unique type I and 6 type II three-dimensional structures of inorganic pyrophosphatase exist within the Protein Data Bank (PDB), with 16 E-PPase and 11 Y-PPase structures available. These two enzymes are the most characterised of all pyrophosphatases. Three structures exist from *Helobacter pylori* and *Bacillus subtilis* (type II), two from *Streptomyces gordonii* (type II) and *Mycobacterium tuberculosis*, and one each from: *Pyrococcus furiosus*, *Pyrococcus horikoshii*, *Streptomyces mutans* (type II), *Sulfolobus acidocaldarius*, and *Thermus thermophilus*. To date, no mammalian PPase structure has been solved. See **table 1.5** for examples of PPase structures deposited in the PDB.

While no structures exist with both magnesium and phosphate bound, magnesium-manganese-sulphate (phosphate analog) (1I74), manganese-phosphate (1WGJ, 1YPP, 8PRK, 1E6A, 1E9G), manganese-sulphate (1WPN, 1K20), zinc-sulphate (1WPP), cobalt-phosphate (1M38), magnesium (1OBW, 1IPW, 1HUJ, 1HUK, 1QEZ) manganese (1INO, 1WGI, 1K23), and sulphate-bound structures (1JFD, 1MJW, 1MJX, 1WPM, 2PRD) further develop our knowledge of the structural basis of PPase catalysis.

1.2.7a Primary structure

As described previously, type I and II PPases are evolutionarily distinct and share no sequence homology, so will not be compared here. Based on primary structure alone, Sivula *et al.* (1999) divided type I PPases into three subfamilies: prokaryotic, with 191 ± 29 residues; plant, 214 ± 3; and animal/fungal, 286 ± 6. Sequence identity between prokaryotic PPases ranges from as low as 23 % to a virtually indistinguishable 99 %. Conservation within the animal/fungal group is also diverse, with members sharing between 42 and 95 % identity. Plant PPases on the other hand share between 74 and 90 % identity, representing the least divergent group.

PDB ID	**Species**	**PPase type**	**Reference**	**Res. (Å)**	**Space group**	**Ligand**
1WPM	*Bacillus subtilis*	II	Fabrichniy *et al.*, 2004	2.05	$P2_12_12_1$	SO_4
1FAJ	*Escherichia coli*	I	Kankare *et al.*, 1996	2.15	H32	None
1. 1I6T 2. 1N40	*Escherichia coli*	I	Samygina *et al.*, 2001	1. 1.20 2. 1.10	H32	1. Ca_3PPi 2. Ca_3
1TWL	*Pyrococcus furiosus*	I	Zhou *et al.*, to be published	2.20	H32	None
1UDE	*Pyrococcus horikoshii*	I	Liu *et al.*, 2004	2.66	$P2_12_12$	None
1PYP	*Saccharomyces cerevisiae*	I	Harutyunyan *et al.*, 1983	3.00	P1121	None
1YPP	*Saccharomyces cerevisiae*	I	Harutyunyan *et al.*, 1996	2.40	$P2_12_12_1$	$(Mn_2Pi)_2$
1. 1WGJ 2. IWGI	*Saccharomyces cerevisiae*	I	Heikinheimo *et al.*, 1996	1. 2.00 2. 2.20	$P2_12_12_1$	1. $(MnPi)_2$ 2. Mn_2
1K20	*Streptomyces gordonii*	II	Ahn *et al.*, 2001	1.50	$P2_12_12_1$	$MnSO_4$
1QEZ	*Sulfolobus acidocaldarius*	I	Leppanen *et al.*, 1999	2.70	P21 (P1211)	Mg

Table 1.5: Examples of inorganic pyrophosphatase structures available in the Protein Data Bank (PDB), as of November 2006.

Plant and prokaryotic PPases share between 27 and 49 % sequence identity and exhibit the same type of deletions. However, inter-group similarity between the plant and animal/fungal groups is below 29 %.

Several insertions exist within animal/fungal PPases which have been implicated in weak membrane association (Vihinen and Lundin, 1992 and Sivula *et al.*, 1999). These additional residues are located between residues Asp36-Arg37 (both conserved active site residues), Met43-Ala44, Asp58-Asp59 (the beginning of a highly conserved region), and Gly116-Ala120 (all Mtb-PPase numbering). Insertions in these regions, with the exception of Gly116-Ala120, occur within Y-PPase (**figure 5.12**, section 5.8.4a) and the Asp58-Asp59 insertion is also found in the prokaryotic *Chlamydia trachomatis*.

An alignment of 37 type I PPases identified 17 conserved residues (Sivula *et al.*, 1999), which were observed by X-ray crystallography to form the active site (Terzyan *et al.*, 1984). A number of structural (Harutyunyan *et al.*, 1996 and Heikinheimo *et al.*, 1996) and biochemical (Lahti and Kolakowski *et al.*, 1990 and Salminen *et al.*, 1995) studies have found 13 of these to be involved in metal/substrate binding (E15, K23, E25, R37, Y49, D59, D61, D64, D91, D96, K98, Y133, and K134, Mtb-PPase numbering). Site-directed mutagenesis studies identified essential residues in E-PPase: Asp97 (Asp91 in Mtb-PPase) and Glu97 (Glu92), to be important in maintaining the structural integrity of the enzyme; and Asp102 (Asp96) and Lys104 (Lys98), to be essential for PPase catalytic activity (Lahti and Pohjanoksa *et al.*, 1990, Efimova *et al.*, 1999, and Hyytia *et al.*, 2001).

Residues found in a number of sequences, but which are not explicitly conserved, include Tyr45, Pro62, Gly76, Phe130, and Lys140 (Sivula *et al.*, 1999). Residues involved in intersubunit interactions are also generally well conserved between the subgroups. The inter-trimeric interactions within prokaryotic PPases are often formed from residues homologous to His128, His132, and Asp135 (Mtb-PPase numbering), as found in 74, 35, and 44 % of the 23 aligned sequences, respectively (Sivula *et al.*, 1999). His132 and Asp135 are substituted with a threonine and an alanine, respectively in T-PPase, and an arginine and a glutamic acid in Pfu- and Pho-PPases. Asp135 is also substituted with a glutamic acid in S-PPase. Hydrophobic interactions which stabilise the trimer however, are poorly conserved. Only three residues involved in interface interactions within the animal/fungal group were found in all of the 9 sequences aligned: Arg51, Trp52, and Trp279 (Y-PPase numbering). Unsurprisingly, none of these interface residues are conserved outside of the groups for which they belong.

1.2.7b Overall fold of type I pyrophosphatases

Whilst prokaryotic PPases and Y-PPase differ in molecular weight and share low sequence identity, the overall fold remains conserved (**figure 1.2**). Y-PPase forms extensions at either end of this core, with a 27-residue N-terminal and 59-residue C-terminal protrusions (Kankare *et al.*, 1994 and Heikinheimo *et al.*, 1996). PPases are predominantly β-structures, with the central core formed from five antiparallel β-strands (β1 and β4-7), which twist into a β-barrel (Teplyakov *et al.*, 1994 and Heikinheimo *et al.*, 1996). The only two major α-helices, together with a long β-hairpin, flank either end of the barrel. One of these helices, α2, forms a lid over the base of the β-barrel (Teplyakov *et al.*, 1994) and the second helix, α1, forms an essential part of the active site cavity wall (Kankare *et al.*, 1996).

1.2.7c Oligomeric state of type I pyrophosphatases

All known type I prokaryotic PPases form ~ 120 kDa hexamers (Teplyakov *et al.*, 1994) under physiological conditions. Leppanen *et al.* (1999) postulated S-PPase to be a symmetric homohexamer in the resting state, however mutations of active site residues which stabilise the hexamer, resulted in dissociation to a dimer of trimers (Salminen *et al.*, 1995).

Intra-trimer interactions are very tight, stabilised by a parallel β-bridge between strands β2-3 (Pho-PPase)/β6 (T-PPase) of one subunit and a β-hairpin (residues Gln71-Val77) of the other (Teplyakov *et al.*, 1994 and Liu *et al.*, 2004), however specific residues are not generally conserved within these regions. Extensive hydrophobic interactions also form between the subunits, stabilising the trimeric structure even further. Intra-hexameric contacts are predominantly formed by symmetry-related helices, α1 (Teplyakov *et al.*, 1994). In E-PPase, stability arises through hydrogen bonds between three α1 residues from each monomer (Baykov *et al.*, 1995 and Kankare *et al.*, 1996).

Figure 1.2: Superimposition of E-PPase 1I6T (Samygina *et al.*, 2001) (olive green) with the central core (residues 41-230) of Y-PPase 1WGJ (Heikinheimo *et al.*, 1996) (dark green). Figure generated using PYMOL (DeLano Scientific).

The oligomeric state of Y-PPase differs from that described for prokaryotes, forming a physiologically active homodimer, with the two active sites ~ 40 Å apart (Harutyunyan *et al.*, 1996). The two monomers are stabilised by the stacking of aromatic rings, with two central histidines hydrogen-bonded to each other (Heikinheimo *et al.*, 1996), however these contacts are relatively loose (Harutyunyan *et al.*, 1996). No active site residues participate in dimer stabilisation, unlike in E-PPase (Heikinheimo *et al.*, 1996).

1.2.7d Active site of type I pyrophosphatases

A cavity between the β-barrel and α1 forms the active site within PPases and is lined with polar residues, with a hydrophobic base (Harutyunyan *et al.*, 1996). The 13 conserved residues mentioned in section 1.2.7a fill the active site and participate in substrate/product and activating metal coordination. In the proposed scheme, described in **figure 1.3**, four metal ions bind to the active site, together with the PPi substrate.

Five conserved, positively charged residues, directly participate in coordination with the activating metals, namely four aspartic acids (Asp59, 64, 91, and 96 Mtb-PPase numbering) and a glutamic acid (Glu25) (Harutyunyan *et al.*, 1996, Heikinheimo *et al.*, 1996, and Harutyunyan *et al.*, 1997). Additional conserved residues indirectly interact with the ions, via water molecules. These include two glutamic acids (Glu15 and 25) and one aspartic acid (Asp61). A number of polar and negatively charged residues, which also form part of the active site, anchor substrate in the correct orientation for hydrolysis. These are Lys23, Arg37, Tyr133, and Lys134 (Heikinheino *et al.*, 1996 and Samygina *et al.*, 2001).

The activating metal binding site designated Mn1 (**figure 1.1**), binds with the highest affinity via interactions with oxygens from three of the aspartic acids (Asp59, 64, and 96), together with two water molecules, one being the nucleophilic hydroxide (section 1.2.5), and a P2 oxygen (Harutyunyan *et al.*, 1996 and Heikinheimo *et al.*, 1996). The association of activating metal site 2, Mn2, is much weaker, with the only direct enzyme ligand being Asp64, together with four water ligands. Binding of the last two metals, Mn3-4, are reasonably similar. Mn3 interacts with the protein through contacts with a Glu25 oxygen and forms additional bonds with three water molecules and two phosphoryl oxygens (P1 and P2). Mn4 makes two contacts with two aspartic acids oxygens, Asp91 and Asp96. This is further stabilised by two water molecules and two phosphoryl oxygens (P1 and P2) (Harutyunyan *et al.*, 1996 and Heikinheimo *et al.*, 1996).

Of the two phosphoryl groups, P1, represents the highest affinity site with its three free oxygens directly coordinated to three PPase residues: Arg37, Tyr133, and Lys134 (Heikinehimo *et al.*, 1996). P1 is further stabilised by two water molecules and a direct attachment to Mn3. P2 on the other hand, is only directly coordinated with one protein residue, Lys23, but interacts with all metal ions, although Mn2 is coordinated via the nucleophilic water molecule. P2 is also bound to an additional two water molecules.

1.2.8 Structure of type II pyrophosphatases

Relatively little structural information is known about type II PPases, however of the six structures solved to date (section 5.8.5), all exhibit a similar fold (Fabrichniy *et al.*, 2004). The structurally characterised PPases from *Streptococcus mutans* (Sm-PPase), *Streptococcus gordonii* (Sg-PPase), and *Bacillus subtilis* (Bs-PPase), are 310 ± 1 residues in length, with a monomeric structure formed of two domains (**figure 1.3**). The N-terminal domain is the largest of the two, comprised of residues 1 – 189 in Sm-PPase, with the smaller C-terminal domain (residues 196 – 309) connected via a linker sequence (residues 190 – 195) (Merckel *et al.*, 2001). The N-terminal domain is formed from a five-stranded parallel β-sheet and seven α-helices, with one helix (αG) immediately preceding the linker sequence which enters the C-terminal, via a further helix (αH). The C-terminal domain is formed of a five-stranded mixed β-sheet, together with three α-helices (Merckel *et al.*, 2001). Type II PPases form physiological dimers, with the interface formed by residues 99 – 115 in Sm-PPase.

Figure 1.3: Cartoon representation of the type II PPase from *Streptococcus mutans* 1I74 (Merckel *et al.*, 2001). The N-terminal domain is shown in dark blue linked (deep pink) to the smaller C-terminal domain, represented in light blue. Figure generated using PYMOL (DeLano Scientific).

The type II active site is located within the domain interface and sequence identity within this region is understandably high (Merckel *et al.*, 2001 and Fabrichniy *et al.*, 2004). Nine of the 36 residues explicitly conserved within the type II family appear to directly participate in metal ion and substrate binding (Merckel *et al.*, 2001), and are described further here. Within the Sm-PPase structure, two manganese ions (M1-2) and one magnesium (M3) were modelled, together with two sulphate molecules (Merckel *et al.*, 2001). Coordination of the two manganese ions occurs via four aspartic acids (Asp12, 14, 75, and 149) and two histidines (His8 and 97). The magnesium does not coordinate directly with any protein residue, highlighting the preference for manganese over magnesium in type II enzymes. This preference is thought to be due to the presence of histidines within the active site, which are generally not found within type I

PPases. Whilst magnesium binds almost exclusively to oxygen ligands, manganese is able to bind to the side-chain nitrogens of histidine (Merckel *et al.*, 2001). The two sulphate molecules (S1-2) in Sm-PPase are bound to the positively charged side chains of Lys205 and Arg295 and S1 is also coordinated to His98.

Despite the complete lack of sequence homology between types I and II PPases, the arrangement of active site residues and their interactions with the metal ions/substrate are surprisingly similar, suggesting an analogous mechanism resultant of convergent evolution (Merckel *et al.*, 2001). The Sm-PPase structure identified a water molecule bridging the two metal sites (M1-2), which they propose acts as the nucleophilic hydroxide due to analogy with Y-PPase (Heikinheimo *et al.*, 1996 and 2001 and Merckel *et al.*, 2001).

1.3 Cytochrome P450s

1.3.1 Haem proteins and cytochromes

Haem proteins perform a vast array of functions across all species, ranging from catalysis (catalases, cytochrome P450s, peroxidases), to electron transfer (cytochromes), oxygen transport and storage (globins), and nitric oxide transport (nitrophorin) (http://metallo.scripps.edu/promise/HAEMMAIN.htm). Of the catalytic haem-containing enzymes, catalase converts hydrogen peroxide, a toxic product of metabolism, to water and oxygen and is implicated in ethanol metabolism, inflammation, apoptosis, ageing, and cancer (Putnam *et al.*, 2000). It has one of the highest turnover rates of all known enzymes, with a conversion rate of 83, 000 molecules per second. Haem proteins with no enzymatic function include vertebrate myoglobin and haemoglobin, which store (myoglobin) and transport (haemoglobin) oxygen within muscle and blood cells, respectively (Voet *et al.*, 1999). Cytochromes are ubiquitous proteins, present in virtually every organism, with the exception of a few obligate anaerobes (Voet *et al.*, 1999). These proteins exploit their ability to alternate between haem iron oxidation states (reduced, Fe^{2+} and oxidised, Fe^{3+}), to enable electron transport.

Each haemoprotein requires a prosthetic haem group, of which nine exist (*a* – *d*, d_1, *o*, P460, and sirohaem), which coordinate to the protein via a central iron atom. Haem type-*a* exists within cytochrome *c* oxidases, whilst the *b*-type is found within *b*-type cytochromes, cytochrome P450s, and some globins and catalases. This type of haem (*b*) is known as a protoporphyrin IX, and is shown in **figure 1.6**. Type-*a* haems differ from protoporphyrin IX in that they contain a long, hydrophobic tail of isoprene units (Voet *et al.*, 1999). Proteins such as *c*-type cytochromes, contain a haem group with cysteine sulfhydryls, which form thioether linkages to the protein.

As well as the haem group with which they bind, cytochromes can also be classified by their haem iron coordination. In type-*a* and –*b* cytochromes, the haem iron is sixth-coordinated to the four porphyrin nitrogens, together with two histidine residues. In type-*c* cytochromes, one histidine is substituted with the sulphur atom from methionine. Finally, in cytochromes P450, one axial ligand is provided by a deprotonated sulphur from a cysteine residue, with the final position able to bind to a number of compounds, such as water, dioxygen, carbon dioxide, and other inhibitors.

1.3.2 Cytochrome P450s

Cytochrome P450s are a superfamily of haem-thiolate proteins that are involved in "phase I" metabolism, whereby a wide array of hydrophobic substrates are monooxygenated (**equation 1.2**) (Nebert and Gonzalez, 1987). Such processes usually produce unstable products which are then further metabolised. P450s generally function as part of electron transport chains, acting as terminal oxidases during processes such as steroid metabolism, drug deactivation, procarcinogen activation, fatty acid metabolism, xenobiotic detoxification, and catabolism of exogenous compounds (Hasemann *et al.*, 1995).

Equation 1.2:

$$\mathrm{SH + O_2 + NAD(P)H + H^+ \longrightarrow SOH + NAD(P)^+ + H_2O}$$

Where *SH* is the substrate and *SOH* is the monooxygenated product.

P450s are so named because in the presence of CO, ferrous enzymes exhibit an intense Soret peak at 450 nm (see section 1.3.8) (Omura and Sato, 1964). These ubiquitous enzymes have been identified in bacteria, fungi, plants, insects, and vertebrates (Nelson *et al.*, 1996). As of 20th October 2006, 6, 422 unique P450 genes have been identified within 708 families. Of these families, 99 are from animals (2, 279 genes), 94 from plants (2, 311 genes), 282 from fungi (1, 001 genes), 177 from bacteria (621 genes), 51 from protists (210 genes), and 5 from archaea (8 genes) (http://drnelson.utmem.edu/CytochromeP450.html). New sequences are added regularly due to ongoing genome projects.

Classification of P450s depends on the electron transfer system utilised, with class I proteins requiring an FAD-containing NAD(P)H ferredoxin reductase, together with an iron-sulphur (redoxin) protein which mediates electron transfer between the reductase and the P450 (Shimizu *et al.*, 2000 and Li, 2001). In contrast, class II P450s require only one redox partner, an FAD/FMN-containing NADPH flavoprotein. Class III enzymes obtain an electron source from an endogenous endoperoxide or hydroperoxide encoded by the same polypeptide, while class IV enzymes receive electrons directly from NAD(P)H (Shimizu *et al.*, 2000). Most bacterial and mitochondrial P450s belong to class I, whilst some bacterial, together with microsomal and fungal P450s, belong to class II.

Plant P450s can be further categorised by their reaction mechanisms: A-type are involved purely in plant-specific biochemical pathways; and B-type reactions, which are more closely related to non-plant P450s (Durst and Nelson, 1995). These enzymes are thought to be involved in highly conserved reactions such as sterol biosynthesis (Morikawa *et al.*, 2006).

1.3.3 Nomenclature

The current nomenclature system for P450s was developed by Nebert and Nelson *et al.* (Nebert and Nelson, 1991, Nebert *et al.*, 1991, and Nelson *et al.*, 1993 and 1996). Cytochrome P450 is represented as "CYP" and is followed by an Arabic number which denotes the family name. When more than one subfamily exists, this is followed by a letter, and finally by an Arabic numeral which represents the individual gene (such as CYP3A4). Where a family has only one member, a subfamily letter and gene number are not always included (such as CYP125). The criteria grouping P450s into families is generally sequence dependent, with those that share > 40 % amino acid sequence identity belonging to the same family. If sequence identity is > 55 %, P450s belong to the same subfamily. However there are some exceptions, as is the case for a number of plant P450s which are classified differently due to gene duplications and shuffling (Werck-Reichart *et al.*, Nielsen and Moller, 2005).

The nomenclature system is organised such that: CYP1 through to CYP69 and CYP301-500 represent animal P450s; CYP71-99 and CYP701-772 represent plants; CYP101-281, bacteria; and CYP501-699, lower eukaryotes. CYP5001 onwards has also been allocated for newly identified animal, fungal, and lower eukaryotic P450s (http://drnelson.utmem.edu/CytochromeP450.html).

1.3.4 The importance of cytochrome P450s

The myriad of metabolic roles performed by these enzymes has generated significant interest worldwide, particularly concerned with the role P450s play during human drug metabolism. Many potential drugs are discarded due to interactions with these enzymes, either they are metabolised too rapidly thus exerting no beneficial effect, or can themselves up or down regulate a P450, thereby affecting the metabolism of another compound (known as drug-drug interactions). Human P450s also function in the oxidation of xenobiotics such as carcinogens, pesticides, steroids, and vitamins (Guengerich, 1995), and can even promote carcinogenesis (Werck-Reichhart and Feyereisen, 2000).

In pathogenic organisms, identification of P450s which are essential for virulence or survival during infection may facilitate novel drug design. A review by Munro *et al.* (2003) proposed the potential of P450s as novel antimycobacterial targets. Twenty-two unique P450s have been identified within the pathogen's genome, the highest number found in any bacterium, suggesting an important role within the organism. Furthermore, elevated levels of P450 have been identified in a number of drug-resistant organisms including *Mycobacterium tuberculosis* (Ramachandran and Gurumurthy, 2002). Specifically, a two-fold increase in P450 content was observed in isoniazid-resistant bacteria, compared with that of the non-resistant strain. They suggest that the importance of certain P450s in the metabolism of two key chemotherapeutics, isoniazid and rifampicin, made increased expression of these enzymes an evolutionary advantage in the presence of these toxic drugs.

Cytochrome P450s account for the largest group of proteins within plants and catalyse the complex regio- and stereospecific-biosynthetic reactions yielding products which enable communication, attract pollinators, and deter pathogens and herbivores (Morant, *et al.,* 2003). They execute critical oxidation steps within plant secondary metabolism, via hydroxylations, dealkylations, dehydrations, and carbon-carbon bond cleavages (Durst and Nelson, 1995). Pathways for which plant P450s play a metabolic role include the phenylpropanoid, terpenoid, and alkaloid pathways, which produce lignins, isoflavonoids, and anthocyanins (Chapple, 1998). These natural products are an attractive target for improving the health and nutritional value of commercial crops and plants, by engineering herbicide resistance or introducing new functions (Feldmann, 2001). A further application of plant P450 studies is in pharmaceuticals, as 25 % of modern medicines are derived from plants and secondary metabolites contribute to many synthetic drugs (Morant *et al.*, 2003).

1.3.5 Mechanism

The cycle begins with an oxidised substrate-free P450 in a low-spin state, with water as the sixth axial iron position (Sligar and Gunsalus, 1976). This is known as the resting state (step 1, **figure 1.4**). Although substrate does not interact directly with the iron (or the haem), its binding does dislodge the 6th axial water (step 2). This dehydration not only causes a spin-shift of the haem iron to a high-spin state (Li, 2001) but has also been shown to increase the redox potential from –300 mV to –170 mV in P450cam (Sligar, 1976). This prevents electron flow from the redox partner (iron-sulphur protein, redox potential approximately -200 mV) to the haem iron in substrate-free systems, by making it thermodynamically unfavourable, thus avoiding unnecessary wastage of reducing equivalents (Mueller *et al.*, 1995 and Li, 2001). The access of water to the 6th axial position has been proven to regulate the haem iron spin state of P450s, alternating between the ferric S = 5/2 (5-coordinated, high-spin system) and S = 1/2 (6-coordinated, low-spin system) (Harris and Lowe, 1993).

In step 3, an electron transferred from the redox partner, reduces the haem iron to the ferrous form (Li, 2001), allowing dioxygen to bind to the 6th position (step 4), forming a LS "oxy-P450" ferric-superoxide species (Tyson *et al.*, 1972 and Sligar *et al.*, 1974). In step 5, a second electron reduces the haem iron to a ferric-dioxo species, which provides a good Lewis base, undergoing protonation to yield a ferric peroxide complex (step 6) (Shaik and De Visser, 2005). Solvent molecules within the active site are thought to provide this source of protons (Poulos and Johnson, 2005), however a conserved threonine residue (Thr252 P450cam numbering) also plays an important role in protonation for some P450s (Shaik and De Visser, 2005). The haem-iron then undergoes a second protonation forming a reactive, high-valent iron-oxo complex, which releases water (step 7). Finally, the distal oxygen is transferred to the substrate and the product is released. Another water coordinates to the 6th axial ligand, bringing the cycle back to step 1 (Shaik and De Visser, 2005).

Figure 1.4: The catalytic mechanism for cytochrome P450s (Shaik and De Visser, 2005). See text for descriptions of stages. The proximal cysteinate ligand is abbreviated as "C" and thick black lines denote the porphyrin. The substrate, SH, is monooxygenated to the product, SOH.

1.3.6 Inhibition

Many compounds interact with P450s and prevent catalysis at various points of the monooxygenation pathway. A review by Correia and Ortiz de Montellano (2005) described the mechanisms employed by a number of common inhibitors, which either bind reversibly to the active site, or (quasi)-irreversibly after the oxidation step (step 4, **figure 1.4**). The latter can often be categorised as "suicide" or mechanism-based inhibitors, whereby a compound only becomes inhibitory after partial or full catalysis by the target enzyme (Voet *et al.*, 1999). Inhibition by mechanism-based methods are highly specific due to the following requirements: the inhibitor must bind to the enzyme initially, then be recognised as a substrate to enable catalytic activation, and finally the reactive species produced must be able to irreversibly alter the enzyme and so stop the cycle (Ortiz de Montellano and Correia, 1995). Knowledge of which compounds inhibit particular P450s and their subsequent mechanism will undoubtedly aid in future drug design.

1.3.6a Reversible inhibition

Compounds which bind to the hydrophobic domain, coordinate to the haem iron, or interact with active site residues can reversibly inhibit P450s (Correia and Ortiz de Montellano, 2005). Different P450 substrates can competitively compete with each other, by binding to the liphophilic regions of the active site. This method of inhibition is not as effective as inhibitors which coordinate to the 6th haem iron position of ferric P450s, via their heteroatomic lone pair electrons. These compounds not only prevent dioxygen binding, but also change the redox potential sufficiently enough to discourage reduction by the P450 reductase partner. Examples of such inhibitors are cyanide (Kitada *et al.*, 1977), NO (Wink *et al.*, 1993), and other hydrophobic nitrogen-containing compounds including pyridine and imidazole derivatives (Testa and Jenner, 1981). The latter two derivatives are potent inhibitors of P450 due to additional strong interactions with liphophilic active site residues.

The importance of these azole-based inhibitors in drug design is well recognized and innumerable publications relating to this field are available. Zhang *et al.* (2002) studied the interactions between various azole-based antifungal agents and human P450s with the aim of predicting potential drug-drug interactions. Of interest, all of the five drugs tested (clotrimazole, miconazole, sulconazole, tioconazole, and ketaconazole) exhibited non-selective inhibition towards the eight P450s studied (CYP1A2, CYP2A6, CYP2C9, CYP2C19, CYP2D6, CYP2B6, CYP2E1, and CYP3A4). Furthermore, another study identified azole compounds as potent inhibitors of mycobacterial P450s (McLean and Marshall *et al.*, 2002).

A number of inhibitors, most notably CO, bind to the 6th axial haem iron position of ferrous P450s. The CO carbon donates electrons to the iron via a σ-bond, in addition to a back-donation of electrons from the occupied iron *d*-orbitals to the empty antibonding π-orbitals of CO (Hanson *et al.*, 1976 and Correia and Ortiz de Montellano, 2005). CO-induced inhibition is relatively weak.

1.3.6b Quasi-irreversible and irreversible inhibition

Mechanism-based inhibition is both highly specific and irreversible, and can affect the P450 in a number of ways. Some sulphur (Dalvi, 1987) and halogenated (Halpert and Neal, 1980) compounds, together with terminal alkyl/aryl olefins and acetylenes (Gan *et al.*, 1984 and Roberts *et al.*, 1993) are catalytically activated by P450 to form a reactive species which covalently binds to the protein (Correia and Ortiz de Montellano, 2005). In some cases they induce an autoimmune response in humans, resulting in destruction of the P450 (Fontana *et al.*, 2005). Terminal olefins and acetylenes also inhibit P450s by covalent bonding, but to the haem group rather than the polypeptide itself (Helvig *et al.*, 1997 and Zhou *et al.*, 2005). In some cases these compounds modify the catalytic activity (Raner *et al.*, 2002).

Another form of mechanism-based inhibition involves the oxidised inhibitor modifying the P450 haem, resulting in an inactive enzyme covalently linked to a degraded haem group (Correia and Ortiz de Montellano, 2005). Compounds which result in such changes include tetrachloromethane (Davies *et al.*, 1986) and spironolactone, a medication used to treat hyperaldosteronism (Osawa and Pohl, 1989).

Finally, methylenedioxy compounds, amines, and 1,1-disubstituted- and acyl-hydrazines, can tightly coordinate to the haem group and inhibit P450s (Ortiz de Montellano and Correia). Such inhibitors are termed quasi-irreversible due to the ability to dislodge them experimentally, for example using lipophilic compounds as in the case for 3,4-methylenedioxyphenyl-1-propene (isosafrole) (Dickins *et al.*, 1979).

1.3.7 Structure

Whilst more than 6, 000 P450 genes have now been sequenced (section 1.3.2), only 169 three-dimensional structures have been deposited within the Protein Data Bank. **Table 1.6** provides examples of P450 structures deposited, as of November 2006. This is predominantly due to complications encountered during the overexpression of soluble protein and during crystallisation, and is particularly evident for eukaryotic P450s which tend to be associated with endoplasmic reticulum or the inner mitochondrial membrane (Wachenfeldt and Johnson, 1995). The first mammalian P450 crystalline structure was determined after the enzyme was engineered to exclude the single N-terminal transmembrane domain (Williams *et al.,* 2000).

Despite these problems, extensive characterisation of P450cam from *Pseudomonas putida*, spanning more than 30 years, drastically improved our knowledge of P450s and the rapid increase in the number of structures being released will further advance this area of research.

PDB ID	**P450**	**Species**	**Reference**	**Res. (Å)**	**Space group**	**Substrate (S) / Inhibitor (I) / Mutant (M)**
1LGF	P450 oxyB	*Amycolatopsis orientalis*	Zerbe *et al.*, 2002	2.20	C2 (C121)	None
1T2B	P450 cin	*Citrobacter braakii*	Meharenna *et al.*, 2004	1.70	P21 (P1211)	1,8-cineole (S)
1EHG	P450 nor	*Fusarium oxysporum*	Shimizu *et al.*, 2000	1.70	$P2_12_12_1$	None
1OG5	CYP 2C9	*Homo sapiens*	Williams *et al.*, 2003	2.55	P321	S-Warfarin (S)
1WOE	CYP 3A4	*Homo sapiens*	Williams *et al.*, 2004	2.80	I222	None
1. 1EA1 2. 1E9X 3. 1U13	CYP51	*Mycobacterium tuberculosis*	Podust *et al.*, 2001	1. 2.21 2. 2.10 3. 2.01	$P2_12_12_1$	1. Flucanazole (I) 2. 4-Phenyl-imidazole (I) 3. C37L/C151T/

						C442A (M)
1. 1H5Z 2. 1X8V	CYP51	*Mycobacterium tuberculosis*	Podust *et al.*, 2004	1. 2.05 2. 1.55	$P2_12_12_1$	1. None 2. Estriol (S)
1. 1N40 2. 1N4G	CYP 121	*Mycobacterium tuberculosis*	Leys *et al.*, 2003	1. 1.06 2. 1.80	P6522	1. None 2. Iodopyrazole (I)
1. 2IJ5 2. 2IJ7	CYP 121	*Mycobacterium tuberculosis*	Seward *et al.*, to be published	1. 1.60 2. 1.90	$P2_12_12_1$	1. None 2. Flucanazole (I)
1DT6	CYP 2C5	*Oryctolagus cuniculus*	Williams *et al.*, 2000	3.00	I222	None
1CPT	P450 terp	*Pseudomonas sp.*	Hasemann *et al.*, 1994	2.30	P6122	None
1Z8O	P450 eryF	*Saccharo-polyspora erythraea*	Nagano *et al.*, 2005	1.70	$P2_12_12_1$	6-Deoxy-erythronolide B (S)
2D0E	CYP 158A2	*Streptomyces coelicolor*	Zhao *et al.*, 2005	2.15	$P2_12_12_1$	2-Hydroxy-naphtho-quinone (S)
2C7X	P450 pikC	*Streptomyces venezuelae*	Sherman *et al.*, 2006	1.75	$P2_12_12_1$	Narbomycin (S)
1IO8	CYP 119	*Sulfolobus solfataricus*	Park *et al.*, 2000	2.00	$P4_32_12$	None

Table 1.6: Examples of cytochrome P450 structures available in the Protein Data Bank (PDB), as of November 2006. Structures of two *Mycobacterium tuberculosis* P450s and their complexes are included.

1.3.7a Primary structure and sequence homology

Of the known P450s, all have molecular weights in the region of 50 to 60 kDa, and comprise of 400 to 530 residues (Chapple, 1998). As described in section 1.3.3, sequence identity between families is extremely low, at less than 15 %. Until recently, three residues were believed to be explicitly conserved throughout the P450 superfamily: Cys357 (P450cam numbering) which provides the proximal thiolate ligand to the haem iron; and Glu287 and Arg290, which form the EXXR motif in the K helix (see section 1.3.7f). This motif was thought to be essential during tertiary folding, however the identification of a new family of P450s, CYP157, that contain a QXXW motif in place of the standard EXXR appears to contradict this (Rupasinghe *et al.*, 2006).

The identification of this novel enzyme demonstrates P450s explicit requirement for only one conserved residue, the proximal cysteinate ligand. Mutations to this residue have been shown to prevent haem incorporation, resulting in a catalytically inactive enzyme (Shimizu *et al.,* 1988). Furthermore, loss of the cysteinate proximal ligand is responsible for the formation of the inactive cytochrome P420 (Perera *et al.*, 2003). P450cam was inactivated by mutating this residue to a histidine, the proximal ligand for another haem-containing protein, cytochrome c-type (Yoshioka *et al.*, 2001). In the presence of CO, the P450cam C357H mutant exhibited a discrete Soret maximum at 420 nm, indicating full conversion to its inactive form, P420 (**figure 1.5**). This was further substantiated by the lack of catalytic activity in the presence of camphor.

Although not explicitly conserved, residues homologous to Thr252 (P450cam numbering) are often found in P450s, forming half of the (E/D)T pair which has been implicated in the mediation of dioxygen activation (Aikens and Sligar, 1994, Tosha *et al.*, 2003, and Nagano *et al.*, 2005). Studies suggest this residue plays a role in oxy-ferrous stabilisation via hydrogen bonds with dioxygen (Gerber and Sligar, 1994). P450s which lack this residue, such as P450eryF which contains an alanine instead, are thought to utilise a water molecule in place of the OH group provided by Thr252 to stabilise the oxy-ferryl (Cupp-Vickery and Poulos, 1995 and Poulos *et al.*, 1995). The crystal structure of CYP121 from *Mycobacterium tuberculosis* identified a serine in place of the standard threonine residue, providing evidence of a second alternate proton delivery pathway. Another important residue found in all P450s which are required to activate molecular oxygen, Phe350 (P450cam numbering), controls the reaction between the haem iron and molecular oxygen (Ost, *et al.,* 2001).

Figure 1.5: Carbon monoxide-complexes of dithionite-reduced P450cam: native (sold line); and C357H mutation of the proximal cysteinate haem-iron ligand (broken line). Figure taken from Yoshioka *et al.*, 2001.

1.3.7b Haem iron coordination

P450s belong to the haem-thiolate group of enzymes and so contain a *b*-type haem (**figure 1.6**) anchored to the protein via the 5th haem iron coordination site and the deprotonated –S^- group of an explicitly conserved cysteinate residue (Mueller *et al.*, 1995). The iron is held within the haem by the four porphyrin nitrogen atoms and has the potential to alternate between a pentacoordinated and a hexacoordinated system by allowing water, CO, NO, or azole compounds to bind to the 6th axial position (**figure 1.6**). This ligand sits in a *trans* position to that of the proximal position (Mueller *et al.*, 1995).

Figure 1.6: Haem structure and iron coordination. **(A)** b-type haem (protoporphyrin IX), **(B)** pentacoordinated haem-thiolate geometry with a cysteinate group as the 5th proximal ligand, & **(C)** hexacoordinated haem-thiolate geometry with dioxygen as the 6th distal ligand. Figures taken from http://metallo.scripps.edu/promise/HAEM_THIOLATE.html.

1.3.7c Conserved structural core

Despite the low sequence identity observed between P450 families, X-ray crystallography has enabled the identification of a common overall fold (**figure 1.7**) (Li, 2001) which to date, remains unique to the P450 superfamily (Poulos and Johnson, 2005). With the distal substrate binding-side of haem facing forwards, and with the N-terminus on the "left" side, P450s resemble a triangular-shaped molecule (Poulos *et al.*, 1995 and Li, 2001). In this orientation the P450 structure can be divided into two regions: a predominantly β-sheet containing domain on the left; and a helical-rich region on the right, with the majority of helices in plane with the haem (**figure 1.7**).

Two long helices (I and L) flank the haem group and form an inner core, surrounded by additional helices from the N-terminal region. Finally, the antiparallel β-sheets form part of the proteins surface (Poulos *et al.*, 1995).

Especially conserved "core" regions are those which surround the haem, comprising of six helices (the D, E, I, & L bundle and helices J & K), together with two sets of β-sheets, and a region known as the 'meander' which forms between the K-helix and the Cys-loop (Graham and Peterson, 1999 and Werck-Reichhart and Feyereisen, 2000).

1.3.7d The Cys-loop

The region containing the proximal thiolate cysteine residue retains a high sequence identity throughout the P450 superfamily and unsurprisingly is also one of the most structurally conserved (Poulos *et al.*, 1995, Werck-Reichhart and Feyereisen, 2000, and Li, 2001). This region comprises Phe350 to Cys357 (P450cam numbering) and forms a β-bulge, similar to an antiparallel β-pair, providing a hydrophobic environment for the cysteine (Hasemann *et al.*, 1995). This arrangement protects the cysteinate ligand, possibly by shielding it from reducing agents within the solvent (Beale and Feinstein, 1976), and also enables it to accept H-bonds from peptide NH groups (Poulos and Johnson, 2005).

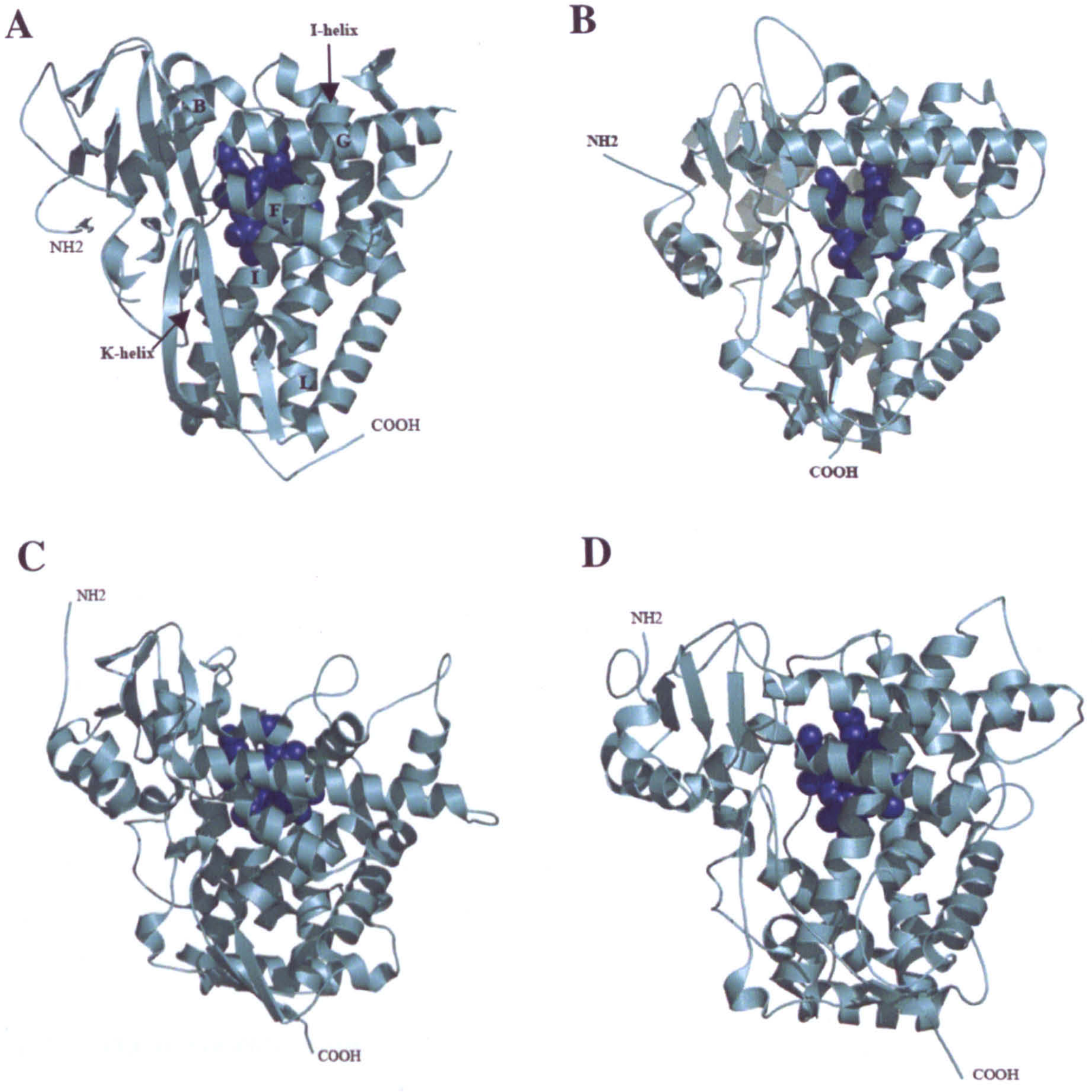


Figure 1.7: Structural comparisons of the overall fold of four cytochrome P450s. **(A)** P450cam 2CPP (Poulos *et al.*, 1987), **(B)** P450eryF 1JIN (Cupp-Vickery *et al.*, 2001), **(C)** CYP51 1EA1 (Podust *et al.*, 2001), and **(D)** CYP2C5 1DT6 (Williams *et al.*, 2000). See text for a generalised description of P450 structure. A number of α-helices, including the I and K helices, are annotated for P450cam **(A)**, however the Cys-loop is not visible from this angle.

1.3.7e I-helix

Another structurally conserved region of P450s is the long I-helix which spans the length of the molecule and helps to form an inner core (**figure 1.7**). This region has been proposed as the central catalytic site (Hasemann *et al.*, 1995) and the (E/D)T motif, conserved in many of these enzymes (see section 1.3.7a), resides in a "kink" which often forms within the I-helix of P450 (Meharenna *et al.*, 2004). This kink occurs due to the donation of a H-bond from the conserved threonine to a carbonyl oxygen within the protein, thus interrupting the helical fold (Poulos and Johnson, 2005). This arrangement is implicated in proton delivery to the oxy-ferryl group in step 7 (**figure 1.4**) (Poulos *et al.*, 1987, Hasemann *et al.*, 1995, and Poulos and Johnson, 2005).

1.3.7f K-helix and the 'meander'

A region known as the 'meander' was first identified in the haemoprotein domain of P450BM-3 (Ravichandran *et al.*, 1993). This region of about 20 residues was so named due to its apparent lack of organized structure, which meandered between the K' helix and the Cys-pocket. Further investigation found near-identical regions in the structures of other P450s, all of which form a specific structure via a hydrogen-bond network between a conserved Arg, His, or Asn residue from the meander, and a highly conserved EXXR motif in the K-helix (Hasemann *et al.*, 1995 and Peterson and Graham-Lorence, 1995). This region has been implicated in the correct binding of haem to P450, and mutations to the K-helix glutamate or arginine have resulted in inactive protein formation (Yoshikawa and Go, 1992 and Hasemann *et al.*, 1995).

1.3.7g Haem coordination

The haem group of P450 is buried within the interior of the enzyme, surrounded by a number of secondary structural elements, namely: the I helix and the N-terminal L helix; the β6-1 and β1-4 strands; the B'-C turn; and the Cys-pocket (Hasemann *et al.*, 1995). Three residues were found to be involved in hydrogen-bonding with the D-ring propionate oxygens via side chain nitrogens in P450terp (Hasemann *et al.*, 1994), and similar configurations have been identified in other P450s including P450cam (Hasemann *et al.*, 1995): His124 and Arg128 (P450terp numbering) are located at the N-terminal of the C helix; and His375 is found within the Cys-pocket. These residues provide the polar and/or charged side chains necessary for propionate coordination within the hydrophobic P450 core (Hasemann *et al.*, 1995). A further six residues participate in an extended hydrogen-bonding network with propionate-bound water molecules in P450terp (Asn72, Phe317 and Arg319, Tyr342, His375, and Trp372), however this configuration is less well conserved within P450s.

1.3.7h Substrate binding region

In contrast to the regions surrounding the haem group, residues involved in substrate recognition are poorly conserved throughout P450s. This reflects the ability of P450s to catalyse a wide range of substrates. The regions involved in substrate binding include helices F and G, which form the substrate access channel entrance in P450BM-3 (Li and Poulos, 2004), and the B' helix which covers the substrate binding pocket (Li, 2001). Modifications to the length of helices F and G, together with alterations in the length of loops flanking the B' helix, enable various substrates to be accommodated within the active sites of different P450s. Such differences are demonstrated by a ~ 90 ° shift between the orientations of the B' helix within the P450cam and P450eryF structures (Poulos and Johnson, 2005).

Characterisation of P450cam identified an eight-fold increase in substrate (camphor) binding in the presence of certain cations (Peterson, 1971 and Mueller *et al.*, 1995). Potassium was later found to bind with the highest affinity (Deprez *et al.*, 1994 and Mueller *et al.*, 1995) and subsequent crystallographic data identified a potential cation binding site at residues Gly93, Glu94, Tyr96, and Glu98 (P450cam numbering) (Peterson, 1971, Poulos *et al.*, 1987, and Mueller *et al.*, 1995), however this has yet to be identified in any other P450 (Li, 2001).

1.3.7i Membrane-binding domains of eukaryotic P450s

Some eukaryotic P450s include a hydrophobic N-terminal helix which anchors to the cytosolic face of the endoplasmic reticulum, co-translationally inserting the enzyme into the membrane. Further signals which target the endoplasmic reticulum have been identified by Szczesna-Skorupa *et al.* (1995), which help to maintain the enzyme's position within the membrane. A number of basic residues also interact with the organelle's membrane lipids, which *in vivo* allows the enzymes to localise where needed. This, together with the N-terminal helix can cause difficulties when attempting to purify and crystallise *in vitro*. Finally, a proline-rich region immediately after the N-terminal helix forms a hinge-like structure, and deletions in this area have been found to disrupt protein structure sufficiently enough to prevent haem incorporation (Szczesna-

Skorupa, *et al.*, 1993 and Yamazaki, *et al.*, 1993). A strategy for the crystallisation of membrane-bound P450s is described by Williams *et al.*, 2000 (see section 1.3.7).

1.3.8 Spectroscopic characterisation

Spectroscopic methods are frequently used to characterise cytochrome P450s. As mentioned in section 1.3.2, these enzymes were named due to their intense absorption at 450 nm in the presence of CO, when in a reduced state (**figures 1.8 – 1.9**). CO binds to the 6th co-ordination site of the haem iron, through electron donation from the carbon to form a σ-bond, competitively inhibiting oxygen binding and thus preventing catalysis (see section 1.3.6a for inhibition mechanism). Such absorption shifts are only observed in haem-thiolate proteins where the thiolate ligand is *trans* to the carbon monoxide molecule (Collman and Sorrell, 1975).

Figure 1.8: Carbon monoxide difference spectra of liver microsomes, taken from Omura and Sato (1964). Spectra from microsomes in the absence of CO were subtracted from the microsomes-CO data. **Curve A:** anaerobic dithionite-reduced microsomes. **Curve B:** aerobic microsomes, in the absence of dithionite. Both curves were recorded in the presence of carbon monoxide.

Figure 1.9: A standard UV/visible spectra of P450BM-3 in different redox/spin states. The wavelength region of 500 to 700 nm has been magnified to illustrate the smaller α and β peaks, to an absorbance range of -0.1 to 0.1. Figure taken from Li *et al.*, 2001.

All P450s exhibit similar absorption spectra, with a Soret peak at approximately 418 nm, with smaller α- and β- bands at ~ 536 and ~ 570 nm, respectively, in the ferric state (**figure 1.9**) (Li, 2001). This Soret band shifts to ~ 408 nm upon reduction of the haem iron. Binding of substrates often induces a blue type-I shift to ~ 390 nm due to displacement of the water from the distal haem-iron position, resulting in a five co-ordinated system (Sligar, 1976, Mueller *et al.*, 1995, 1995 and Li, 2001). Such changes effect the distribution of the outer shell electrons within the haem iron, changing its electronic configuration from S = 1/2 in the low-spin state to S = 5/2 in the high spin.

P450 spin-shifts have also been characterised by electron paramagnetic resonance (EPR). P450cam exhibits g-values of: 2.45 (g_z), 2.26 (g_y), and 1.91 (g_x), for the substrate-free enzyme, which is characteristic of a low-spin haem iron (**figure 1.10A**) (Tsai, *et al.*, 1970 and Lipscomb, 1980). Similar results were also obtained for substrate-free ferric P450BM-3: 2.42, 2.26, 1.96 (Miles *et al.*, 1992); and CYP121 from *Mycobacterium tuberculosis*: 2.48, 2.25, 1.90 (McLean *et al.*, 2005). Upon substrate binding, the system shifts to a predominantly high-spin, with g-values of: 7.85, 3.97, and 1.78 (**figure 1.10B**). Conversely, an EPR spectrum characteristic of a high-spin haem system was obtained for substrate-free HPL (hydroperoxide lyase, CYP74C3) from

Medicago truncatula, with g-values of: 8.03, 3.51, 1.68 (Hughes *et al.*, 2006). This shifted to a low-spin system upon addition of its substrate, 13-HPOTE (13- S-hydroperoxyoctadeca-9Z), with corresponding g-values of: 2.39, 2.24, 1.93. Some inhibitors which bind strongly to the 6th axial haem iron position, such as azole compounds (section 1.3.6a), were found to induce a type-II red shift to approximately 425 – 435 nm (Jefcoate, 1978). This occurs due to the haem-iron adopting a 6-coordination, low-spin configuration.

Figure 1.10: EPR spectra of P450cam (1.1 mM) at 15 °K: **(A)** substrate-free, typical of a low-spin haem iron and **(B)** in the presence of 1.5 mM D-camphor, characterstic of a predominantly high-spin system. g-values are shown. Figure modified from Tsai *et al.*, 1970.

Spectroscopic methods have also been exploited to determine secondary structure within P450s, through the use of circular dichroism (CD). CD measurements of Mtb-CYP121 and the haem-domain of P450BM-3 (both 3 μm), recorded in the far-UV region (190 – 260 nm), identified greater than 50 % α-helical content in both enzymes, **figure 1.11** (McLean and Cheesman *et al.*, 2002).

A study by Yun *et al.* (1996) identified a proportional relationship between salt concentration and helix content of rabbit CYP1A2, measurable by CD (**figure 1.12**).

The helical content was increased from ~ 30 % (in 80 mM potassium phosphate, pH 7.4), to ~ 36 % in the presence of 0.05 M NaCl, and ~ 49 % in 0.1 M NaCl. However, further work by the same group, using rat CYP2B1, did not identify such dramatic changes, with α-helical content increasing by just 5 % in the presence of 0.1 M NaCl (in 50mM potassium phosphate, pH 7.4) to ~ 58 % (Yun *et al.*, 1998).

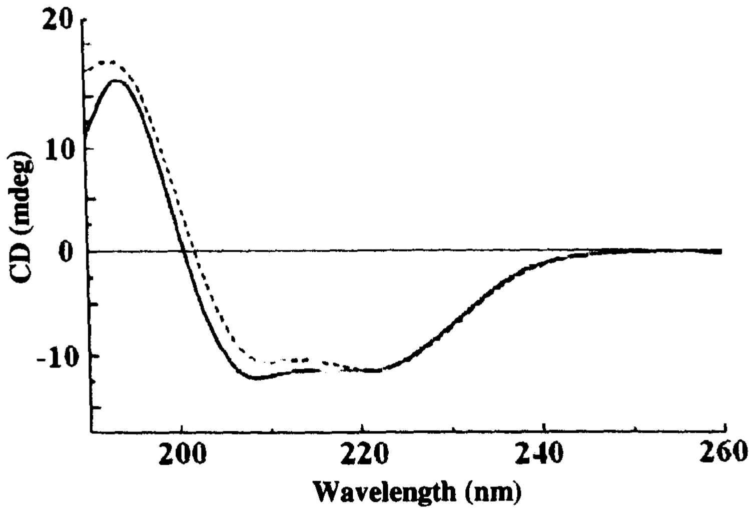


Figure 1.11: CD spectra for Mtb-CYP121 and the haem-domain of P450BM-3 (both 3μm) in the far-UV region (190 to 260 nm). Mtb-CYP121 is represented as a solid line and P450BM-3 (haem-domain) as a broken line. Figure modified from McLean and Cheesman *et al.*, 2002.

Figure 1.12: Effect of ionic strength on the α-helix content of CYP1A2 from *Oryctolagus cuniculus* (rabbit) (Yun *et al.*, 1996). Data were measured in the far-UV region (190 to 260 nm) at a concentration of 1μm. Figure modified from Yun *et al.*, 1996.

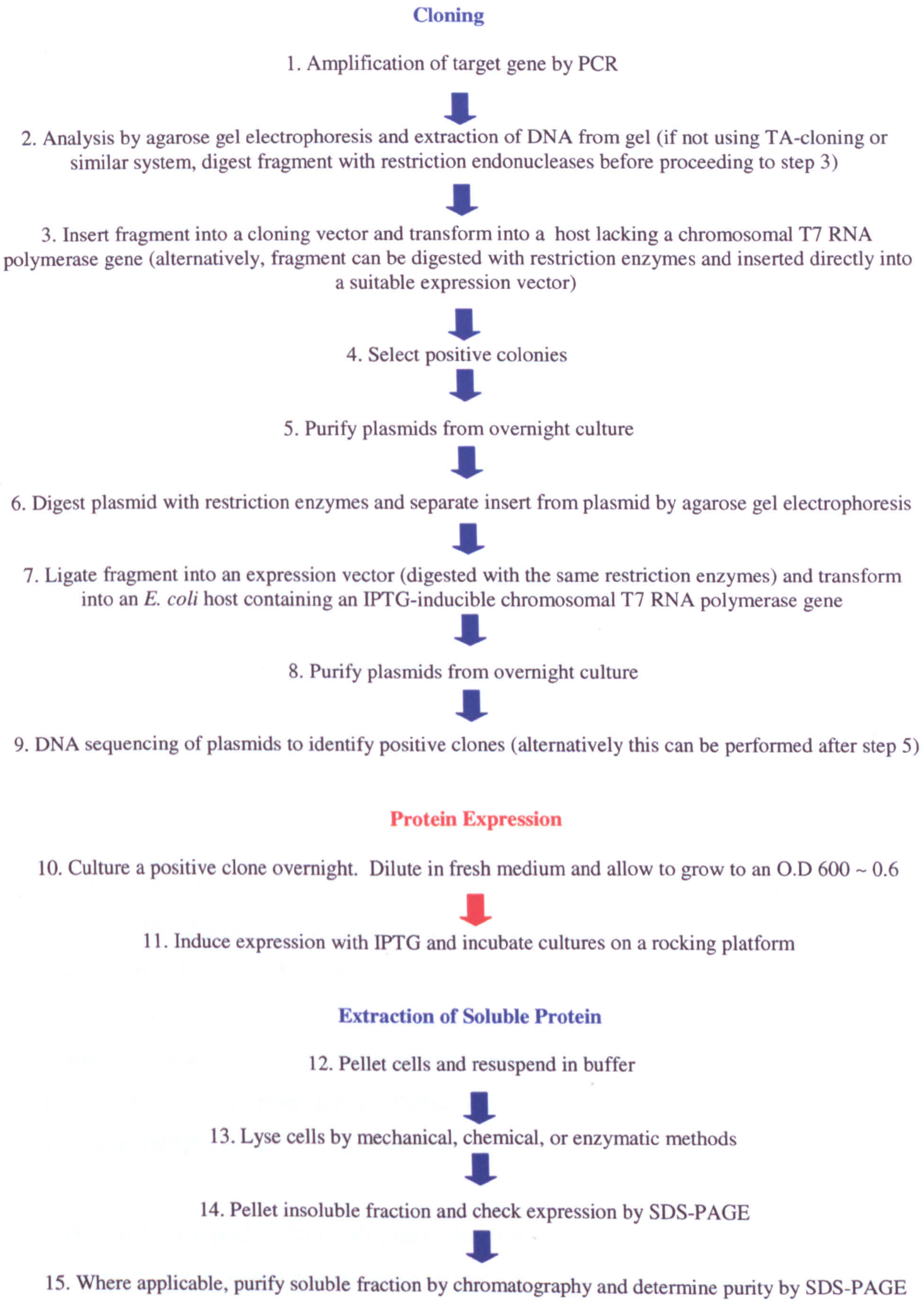
Chapter 2 – Theoretical and experimental background to protein production and characterisation

2.1 *In vivo* protein expression

2.1.1 Introduction

A number of possibilities exist for the production of heterologous proteins *in vivo*, such as the prokaryotic (*Escherichia coli and Bacillus subtilis*) and eukaryotic (*Saccharomyces cerevisiae*, immortalised mammalian cell lines, and *Spodoptera frugiperda* Sf21 cell line) systems. Whilst all have their advantages, bacterial expression systems, and in particular those which use *E. coli* as a host strain, are the most common (Pouwels, 1992 and Sorensen and Mortensen, 2005). Inexpensive cultures can be easily grown overnight and genetic manipulation protocols are well established. Unless the protein of interest is toxic within the host cell, foreign proteins are generally well-tolerated in *E. coli*. Lack of post-translational modification can however be problematic in bacterial systems, for example when expressing eukaryotic proteins which require glycosylation. Another consideration is the possibility of the recombinant protein being expressed as an inclusion body, however this can sometimes be overcome by denaturing the insoluble protein in high salt (such as 6 M guanidine hydrochloride), and then slowly refolding by decreasing the ionic concentration (Whittington, 1989). Recombinant proteins with non-*E. coli* codon usage (such as CGG for arginine and AUA for isoleucine) can successfully be expressed in modified strains such as Rosetta 2 (DE3) by Novagen.

A common *E. coli* procedure, which was used to express recombinant proteins in section 4.3, is described further in this chapter (**figure 2.1**). This prokaryotic system involves the isolation of target DNA from a genomic source using PCR, which is subsequently inserted into a bacterial plasmid vector containing an antibiotic resistance gene. The construct is transformed into an expression strain and positive transformants, selected for by antibiotic resistance screening, are grown in culture medium containing isopropyl-β-D-thiogalactopyranoside (IPTG). This induces transcription of T7 RNA polymerase from the host chromosome which, in turn, transcribes target genes from the recombinant plasmid.



Cloning

1. Amplification of target gene by PCR

2. Analysis by agarose gel electrophoresis and extraction of DNA from gel (if not using TA-cloning or similar system, digest fragment with restriction endonucleases before proceeding to step 3)

3. Insert fragment into a cloning vector and transform into a host lacking a chromosomal T7 RNA polymerase gene (alternatively, fragment can be digested with restriction enzymes and inserted directly into a suitable expression vector)

4. Select positive colonies

5. Purify plasmids from overnight culture

6. Digest plasmid with restriction enzymes and separate insert from plasmid by agarose gel electrophoresis

7. Ligate fragment into an expression vector (digested with the same restriction enzymes) and transform into an *E. coli* host containing an IPTG-inducible chromosomal T7 RNA polymerase gene

8. Purify plasmids from overnight culture

9. DNA sequencing of plasmids to identify positive clones (alternatively this can be performed after step 5)

Protein Expression

10. Culture a positive clone overnight. Dilute in fresh medium and allow to grow to an O.D 600 ~ 0.6

11. Induce expression with IPTG and incubate cultures on a rocking platform

Extraction of Soluble Protein

12. Pellet cells and resuspend in buffer

13. Lyse cells by mechanical, chemical, or enzymatic methods

14. Pellet insoluble fraction and check expression by SDS-PAGE

15. Where applicable, purify soluble fraction by chromatography and determine purity by SDS-PAGE

Figure 2.1: Schematic representation of a common procedure used for the expression of recombinant proteins in a T7 bacterial (*E. coli*) *in vivo* system. This system was used to express proteins in section 4.3.

PAGE
MISSING
IN
ORIGINAL

2.1.2 Polymerase chain reaction (PCR)

Amplification of target genes (step 1, **figure 2.1**) can be performed using the polymerase chain reaction (PCR), first conceived by Kary Mullis in the 1980s (Saiki *et al.*, 1985). This method enables the highly sensitive synthesis of regions of DNA from larger fragments. Target DNA is amplified a million-fold in a matter of hours, without the need for cellular cloning.

A known part of the DNA sequence is used to design two synthetic oligonucleotides (primers), at each end of the region to be amplified. Primers are designed such that each is complementary to one strand of DNA only ("forward" and "reverse" strand primers). Additional sequences may also be added to either end of the final PCR product by engineering the primers with complementary sequences. Such additional elements may include restriction sites, linker sequences, signal peptides, or purification tags.

Target DNA is combined with these oligonucleotide primers, together with a thermostable bacterial DNA polymerase, free deoxy-nucleotides, and a polymerase reaction buffer containing $MgCl_2$. The reaction is first heated to above 94 °C to denature the double stranded DNA, resulting in two single strands, and then cooled to approximately 40 to 60 °C. This allows the hybridisation of oligonucleotide primers to complementary sequences on the target DNA and is known as the annealing step. Precise temperatures for this step require optimisation and are dependent upon the melting point of the primers used.

During the extension period, reactions are heated to approximately 74 °C and the regions of DNA downstream from the synthetic primers are synthesised by DNA polymerase using free dNTP's included in the reaction.

As the cycle is repeated 20 to 40 times, the newly synthesised fragments, together with the primers, act as templates resulting in the rapid synthesis of a single species of DNA fragments. Correctly sized PCR fragments can be identified by agarose gel electrophoresis.

2.1.3 Restriction digestion

Restriction digestion refers to the highly specific cleavage of DNA molecules by endonucleases, resulting in discrete fragments which can be re-ligated to complementary sequences by DNA ligase (Brown, 1992). Such cleavage is necessary when inserting gene fragments into certain vectors (steps 2 and 6, **figure 2.1**). Both the plasmid and the fragment to be inserted are digested with the same enzymes to create complementary ends.

Several hundred of these enzymes have been isolated from prokaryotic sources and are commercially available, allowing for the manipulation of DNA molecules within the laboratory. *In situ*, restriction endonucleases protect bacteria and some viruses from foreign DNA molecules by cleaving them at specific recognition sequences. These enzymes are usually coupled with a modification enzyme, such as DNA-methyltransferase which protects the cells own DNA from cleavage at these sites. A methyl group is added to one base pair of the recognition sequence on each strand, preventing cleavage by the endonuclease. Such restriction-modification systems may be formed of two separate proteins or by two domains in a multi-subunit complex. The type II enzymes used for laboratory purposes cleave within their recognition sequence. Most of these recognise symmetrical DNA sequences and bind as homodimers, however a few bind as heterodimers to asymmetrical sequences. The efficiency of cleavage can be visualised by agarose gel electrophoresis.

2.1.4 Agarose gel electrophoresis

Agarose gel electrophoresis is a method used to separate DNA molecules, predominantly as a function of DNA size and conformation. In the expression protocol described in **figure 2.1**, agarose gel electrophoresis is used to identify PCR products of correct size (step 2) and to separate fragments from restriction digestion (step 6). When agarose, a linear polymer derived from seaweed, is heated in buffer and subsequently cooled, it forms a matrix whose density is proportional to the percentage of agarose. Ethidium bromide, which fluoresces under ultra violet light, intercalates between the bases of the double stranded helix, and is commonly used to stain the gel (Andrews, 1992). Linear molecules become saturated with ethidium bromide, whilst supercoiled bind to a finite number of dye molecules due to the introduction of superhelical turns. This results in linear fragments appearing brighter under U.V light than for supercoiled DNA of the same concentration.

DNA samples are applied to the gel in wells formed by adding a comb to the gel tray before cooling. The negatively charged DNA migrates towards the cathode when an electrical field is applied across the gel. Linear fragments migrate through the gel matrix at a rate inversely proportional to the $\log_{10}$ of the number of base pairs (Helling *et al.*, 1974). Linear, super-coiled, and nicked circular DNA of the same molecular weight migrate at different rates through an agarose matrix, however the rate of migration is determined by the running buffer and electrical current used (Thorne, 1966).

2.1.4a Extraction of DNA from agarose gels

DNA molecules visualised by agarose gel electrophoresis can easily be extracted using commercially available kits, such as the QIAquick® gel extraction kit from Qiagen which incorporates spin columns containing a silica membrane. Gel slices are dissolved in buffer at 55 °C and then applied to the spin column. DNA binds to the membrane in high salt concentrations by adsorption and enzymes, buffers, and other contaminants are removed by washing the column with buffer containing ethanol. Finally, DNA is eluted with water or a low salt buffer.

2.1.5 Cloning of target DNA

Before expression of recombinant protein can proceed, PCR-amplified target DNA must first be inserted into a plasmid vector and transformed into an *E. coli* host lacking the T7 RNA polymerase gene (step 3, **figure 2.1**). This allows for the stable establishment of positive clones (step 4, **figure 2.1**), without expression of recombinant protein. PCR products can either be digested with restriction endonucleases, if such recognition sequences are introduced by primers, and ligated into a vector with complementary ends, or directly inserted into a suitable vector (such as TA-cloning, see section 2.1.5a) without the need for such enzymes. Positive clones can be identified by antibiotic screening, blue/white colony screening, restriction digestion, or DNA sequencing.

2.1.5a Cloning vectors

TA-cloning is a common system used to clone recombinant genes, and is described further here. PCR products synthesised with *Taq* DNA polymerase produce fragments with 3' single deoxyadenosine overhangs, which can be ligated into a linear vector with corresponding 3' terminal thymidines. This reaction is performed by T4 DNA ligase, isolated from *E. coli*, which catalyzes the formation of phosphodiester bonds between neighbouring 3'-hydroxyl and 5'-phosphate ends in double-stranded DNA. As these vectors contain restriction sites in their cloning regions, the introduction of such sites by PCR is not essential. It is sometimes necessary however when specific restriction sites are not present in both the cloning and expression vectors. Restriction sites are not expressed as part of the target gene as the start codon is placed downstream of the 5' site and a termination codon is included before the 3' site.

The pGEM®-T Vector System from Promega enables PCR products to be inserted into the plasmid via TA cloning (**figure 2.2**). Following cleavage of pGEM®-5Zf(+) with *Eco*R V, 3′ terminal thymidines are added, preventing recircularisation. *Taq* polymerase synthesised DNA fragments are ligated into the linearised plasmid by T4 DNA ligase. The multiple cloning region of pGEM®-T exists within an α-peptide coding region for β-galactosidase (*lacZ*), allowing for blue/white colony screening of positive transformants when plated onto LB agar containing IPTG (Isopropyl-β-D-thiogalactopyranoside) and X-gal (5-bromo-4-chloro-3-indolyl-β-D-galacto-pyranoside) (step 4, **figure 2.1**). IPTG induces transcription of the *lacZ* gene, producing β-galactosidase which metabolises X-gal to a blue product, resulting in blue colonies. When genes are cloned in-frame into this region, insertional inactivation prevents the transcription of the lacZ gene and so X-gal, and the colonies, remains colourless.

Figure 2.2: Promega's pGem®-T vector, a commonly used vector for the TA-cloning of genes. Figure reproduced with permission from Promega Corporation.

2.1.5b Cloning hosts

E. coli cells which lack the λDE3 lysogen are suitable for initial cloning as they lack a chromosomal copy of the T7 RNA polymerase gene. A number of commercially available cloning hosts are deficient in both genomic and episomal copies of the lacZ. These cells are thus suitable for blue/white screening of positive clones, when transformed with plasmids containing an α-peptide coding region for β-galactosidase. This is in addition to the standard antibiotic-resistance screening used to select positive transformants (step 4, **figure 2.1**).

An example of a host routinely used for initial cloning steps is Novagen's NovaBlue® cells, which exhibit a high transformation efficiency and are suitable for blue/white colony screening. They confer tetracycline resistance, allowing for additional confirmation of positive colonies.

2.1.5c Transformation of host cells with recombinant plasmids

Whilst a number of procedures exist for the transformation of competent host cells with recombinant plasmids, the heat-shock method first described by Cohen *et al.* in 1972 is both economical and easy to perform. Cells are made competent during the early log phase of growth by washing in ice-cold 0.1 M calcium chloride, however the precise mechanism remains unknown. Supercoiled plasmid DNA is added to the cells on ice and heat-shocked in a water bath at 42 °C before replacing on ice. Cells are grown in an antibiotic-free nutrient rich media, such as SOC, to allow the cells to recover and express the antibiotic resistance gene from the recombinant plasmid. Cells are then plated onto LB agar containing the specific antibiotic, to select for positive transformants.

2.1.5d Purification of plasmid DNA

To enable detection of positive clones by restriction digestion and DNA sequencing, it is first necessary to purify the recombinant plasmid from cell cultures (step 5, **figure 2.1**). Overnight cultures, grown from single transformation colonies, are centrifuged to obtain cell pellets. A commonly used protocol performs an alkaline lysis step to break the cells, and the cleared lysate, obtained by centrifugation, is applied to a silica membrane within a spin column. Plasmid DNA is adsorbed onto the membrane in a high salt buffer and contaminants are removed by washing with buffer containing ethanol. DNA is eluted with a low salt buffer or water. Many commercially available kits are designed for the purification of plasmid DNA, such as Wizard® *Plus* Miniprep (Promega) and QIAprep® Spin Miniprep (Qiagen).

2.1.5e Expression vectors

Target DNA is cleaved from the cloning vector using restriction enzymes and ligated into the multiple cloning site of a vector suitable for expression (steps 6 - 7, **figure 2.1**). Alternately, oligonucleotide primers may be designed to include restriction sites, enabling PCR fragments to be directly inserted into an expression vector. The well established pET system (Plasmids for Expression by T7 RNA polymerase), developed by Studier *et al.* in 1986, provides an efficient construct for the expression of recombinant proteins in *E. coli* hosts under strong control of the bacteriophage T7 promoter. Such systems can direct most of the cells resources to the expression of target protein.

The pET expression system uses host cells which are lysogens of λDE3 and therefore contain a chromosomal copy of T7 RNA polymerase within a *lac* operon, under the control of a *lac*UV5 promoter. This polymerase directs the transcription of target genes at the T7 promoter site on the recombinant plasmid.

Preventing destabilisation of the system, particularly when toxic gene products are incorporated, is controlled with a mechanism based upon lactose operon regulation in *E. coli*. Transcription of T7 RNA polymerase is naturally inhibited by the expression of a *lac* repressor protein, encoded by a chromosomal copy of the *lac*I gene, which binds reversibly to the *lac* operator. This inhibits subsequent transcription of target genes on the recombinant plasmid, although some basal transcription often remains. Such transcription can be reduced further by the addition of a *lac*I gene upstream from the plasmid T7 promoter which prevents the polymerase from synthesising the RNA chain.

Whilst this system provides a highly efficient mechanism for repression, it does not interfere with target gene transcription upon induction with IPTG. IPTG relieves the inhibition by binding to an allosteric site on the repressor protein, causing a conformational change and so decreasing its affinity for the *lac* operator. IPTG is used preferentially over the natural inducer lactose as it cannot be broken down within the cells and so the concentration remains constant. Further control can be achieved by the inclusion of T7 lysozyme which naturally inhibits T7 RNA polymerase activity.

Two commercially available pET vectors, which were used to express recombinant proteins in section 4.3, are briefly described here (**figure 2.3**). The pET-17b vector (Novagen) contains an N-terminal 11aa T7•Tag® sequence followed by a multiple cloning region. Histidine tags are not included in the vector but can be added to target DNA during PCR. pET-28a vectors (Novagen) carry an N-terminal His•Tag® in addition to a thrombin cleavage site and a T7•Tag®. An optional C-terminal His•Tag can be removed from resulting target protein by the addition of a termination codon sequence.

2.1.5f Expression hosts

Expression plasmids are transformed into *E. coli* hosts suitable for protein production and cultured overnight to enable the purification of plasmids (steps 7 - 8, **figure 2.1**). Plasmids are then sequenced (step 9, **figure 2.1**), however this can performed after the initial cloning step instead (after step 5, **figure 2.1**).

Figure 2.3: Examples of Novagen's pET system expression vectors, pET17b and pET28a. Figures reproduced with permission from Merck Chemicals Ltd.

The *E. coli* strain BL21 (DE3), is one of the most common general purpose expression hosts used in laboratories worldwide. Whilst being deficient in both ompT and ion proteases, the DE3 lysogenic strain contains a chromosomal T7 RNA polymerase gene under control of the *lac*UV5 promoter (Novagen, 2006).

However, for expressing recombinant genes which exhibit usage of a high percentage of non-*E. coli* codons (AGA, AGG, AUA, CCC, CGA, CGG, CUA, GGA, and UUA) the BL21 derivative, Rosetta 2 (DE3), can be used to provide a universal translation system (Brinkmann *et al.*, 1989, Seidel *et al.*, 1992, Del Tito *et al.*, 1995, and Rosenburg, 1996). These Novagen cells contain a chloramphenicol-resistant pRARE plasmid which encodes tRNAs for these unusual amino acid codons, under control of their native promoters (Novy *et al.*, 2001). This improves the success rate of expression of such ORFs (open reading frames). Another *E. coli* expression strain, HMS174 (DE3), includes a *recA* mutation in a K-12 background, which can stabilize certain target genes whose products cause the loss of the DE3 prophage (Novagen, 2006).

2.1.6 DNA sequencing

Determining the precise nucleotide sequence of a cloned fragment is an essential step in molecular cloning (step 9, **figure 2.1**). If the target gene is already known, sequencing verifies that the correct fragment has been synthesised and identifies any mutations which may have arisen during amplification. The method also distinguishes genes which are correctly cloned in-frame with the start codon from those which are not. Recombinant plasmids are transformed into an *E. coli* host and grown overnight in LB media. Plasmids are purified as outlined in section 2.1.5d, in preparation for the sequencing reaction.

A number of sequencing methods have been developed over the last 40 years, the most popular being the chain termination procedure described by Sanger in 1977 (Sanger *et al.*, 1977). A variation of this method, dye terminator sequencing, is widely used and will be described here. The reaction includes template DNA, oligonucleotide primers complementary to a region of the template DNA which form the start point of amplification, DNA polymerase, the four deoxynucleotide bases (dATP, dGTP, dCTP, and dTTP), and a low concentration of four dideoxynucleotide chain-terminators. In the case of molecular cloning, the primers flank either side of the gene of interest and it is commonplace to use "universal" primers for regions contained on the plasmid vector such as the T7 promoter and T7 terminator sequences. Dideoxynucleotides lack the 3'-OH group essential for chain extension and are each labelled with dyes which fluoresce at different wavelengths. As the DNA chain is replicated, the dideoxynucleotides are incorporated at random, thereby terminating the sequence and resulting in many related DNA fragments of varying length. Fragments are separated by size on a polyacrylamide gel and fluorescence at different wavelengths is detected. The autoradiogram output shows peaks of different colours, each representing a different dideoxynucleotide. From this, the nucleotide sequence can be inferred by the largest peak at each point (**figure 2.4**).

Figure 2.4: Output file from Sanger dye terminator sequencing. Each colour represents a different dideoxynucleotide, the order of which represent the DNA sequence. Figure taken from www.wikipedia.org.

2.1.7 Expression of recombinant protein

Positive transformants of an expression vector, selected for by antibiotic resistance screening, restriction digestion and DNA sequencing, are grown on a small scale in a suitable medium overnight. Luria-Bertani broth (LB) is a standard media used for this purpose (see appendix 2) (step 10, **figure 2.1**).

Overnight culture is diluted in fresh medium and allowed to grow at 37 °C until the optical density at 600 nm reaches 0.6 - 0.8 (mid-log phase) and then expression of recombinant protein is induced by the addition of ITPG (step 11, **figure 2.1**). If extended incubations, above 24 hours, are required for expression, it is necessary to use a nutrient-rich media such as terrific broth (TB) (see appendix 2).

Small-scale cultures are performed initially to determine the optimum incubation parameters such as temperature, time, and inducer concentrations, before scaling up. Multiple litres of culture may be grown from several millilitres of overnight culture, and can result in milligram quantities of recombinant protein being produced.

2.1.8 Extraction of protein

Cells are pelleted by centrifugation and resuspended in buffer before lysis using mechanical, chemical, or enzymatic means. French pressure cells mechanically lyse cells by forcing crude slurries through a tight space at very high pressures of around 10, 000 psi (pounds per square inch). The sudden release of pressure as the sample is released causes the cells to burst open (Salusbury, 1992). Another mechanical method, sonication, applies frequencies of over 20 kHz to the sample, resulting in the production of gas bubbles. When these collapse, shock waves are formed which lyse the cells (Salusbury, 1992).

Chemical lysis generally incorporates the use of detergents and may include solvents which stimulate autolysis (Goodwin, 1992). A commercially available chemical lysis method is the BugBuster® Protein Extraction Reagent (Novagen) which gently disrupts *E. coli* cells through a combination of detergents. Enzymatic disruption is generally gentler than the mechanical or chemical methods mentioned previously. Enzymes such as trypsin, lysozyme, and other proteases disrupt the cell wall, with full lysis completed by osmotic shock or gentle mechanical treatment (Goodwin, 1992).

The French pressure cell method is generally very successful at lysing bacterial cells and so was predominantly used throughout the work described in this thesis. Soluble fractions are obtained by centrifugation and the extent of expression is determined by SDS-PAGE. When sufficient target protein exists within the soluble fraction, soluble protein extracts can then be purified by a number of chromatographic steps. Expression conditions can be further optimised to obtain greater yields of soluble protein. A number of techniques exist to overcome the production of target protein as inclusion bodies, one of which is described in section 2.1.1.

2.1.9 Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE)

Proteins can be separated, as a function of size, using sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) as first described by Laemmli in 1970. In the presence of an electrical field, protein molecules negatively charged by the binding of SDS, reduced by 2-mercaptoethanol, and denatured at 100 °C, migrate towards the positively charged anode. Anionic SDS detergent binds to the polypeptide backbone at a constant molecular ratio of 1.4 g of SDS : 1 g of protein (Reynolds and Tanford, 1970), conferring a net negative charge to the molecule. The reducing agent, 2-mercaptoethanol, is added to reduce disulphide bridges allowing the protein to adopt a random-coil configuration.

The discontinuous system, developed by Laemmli (1970) based upon work by Ornstein and Davis (both 1964), consists of a large-pore gel which stacks and concentrates samples before they progress into a second more-restrictive resolving gel. This greatly improves the resolution at which proteins are separated compared with continuous buffer systems.

2.2 Cell-free protein expression

2.2.1 A brief history

Cell-free expression systems were developed as an alternative to the traditional *in vivo* methods and to overcome problems associated with the use of living cells. Cell-free systems offer advantages over *in vivo* methods, such as the ability to express host-toxic proteins and the high-throughput manner in which expression can be performed (Katzen *et al.*, 2005).

Early cell-free systems for protein synthesis were based on cytoplasmic animal cell extracts free from mitochondria (Littlefield *et al.*, 1955) and later, from bacterial extracts (Schachtschabel and Zillig, 1959; Lamborg and Zamecnik, 1960; and Tissiéres *et al.*, 1960), both of which expressed endogenous mRNA's only. The first bacterial system to allow for the translation of exogenous mRNA's was developed in 1961 by Nirenberg and Matthaei. Endogenous mRNA's were removed by incubating the cell extract at physiological temperature, allowing ribosomes to accept exogenous templates.

Eukaryotic cell-free systems were developed to allow for the translation of exogenous mRNA by treating rabbit teiculocytes with micrococcal Ca^{2+}-dependent RNase (Pelham and Jackson, 1976). Wheat germ extracts could be used directly for translation due to low level endogenous mRNA (Roberts and Paterson, 1973 and Marcus *et al.*, 1974). The bacterial cell-free system was further developed to allow for a coupled transcription-translation reaction, whereby translation occurs whilst mRNA is synthesised from template DNA by an endogenous RNA polymerase (Lederman and Zubay, 1967).

2.2.2 Composition of cell-free systems

Modern bacterial cell-free reaction solutions contain all of the components required for transcription and translation of target proteins. Whilst the composition of such systems often require optimisation, dependent upon individual protein characteristics, they are usually based upon one of two crude cell extracts including: 1. ribosomes and all soluble enzymes, translation factors, and tRNAs, known as an *E. coli* S30 extract; 2. a combination of ribosome-free extract (S100 extract) plus isolated ribosomes. In the case of codon bias or unusual codon usage, the composition of individual tRNAs can be adjusted (Chumpolkulwong *et al.*, 2006). Nucleotide tri-phosphates (NTPs) provide an essential energy source for cell-free translation, however their role is finite due to NTP-dependent metabolic reactions and the presence of NTPases in the bacterial extract.

Continuous cell-free expression systems overcame this problem by providing a continual supply of essential components such as amino acids and NTPs whilst simultaneously removing reaction waste products (Baranov *et al.*, 1989 and 2002). This was achieved by dialysing the cell-free reaction against a larger feeding solution, across a membrane with a molecular weight lower than that of the target protein and the protein-synthesising machinery (**figure 2.5**). Further optimisation of the dialysis system has allowed members of the Protein Research Group at the RIKEN Yokohama Institute to achieve up to 8 mg of target protein per millilitre of cell-free reaction solution (Kigawa *et al.*, 1999, 2002, and 2004 and Yokoyama, 2003).

Briefly, target genes are amplified by PCR (see section 2.1.2), ligated into cloning vectors and transformed into a host cell lacking a chromosomal copy of the T7 RNA polymerase gene (see section 2.1.5c). Plasmids are purified from the culture as outlined in section 2.1.5d. Once positive clones have been identified by DNA sequencing (see section 2.1.6), they are included directly in the cell-free reaction solution as the DNA template for transcription (**figure 2.6**).

Figure 2.5: A representation of the *E. coli* cell-free system used to express recombinant proteins in section 4.2. All components essential for transcription, translation, and ATP regeneration, together with the target DNA template, are included within the internal "reaction" solution. A dialysis membrane simultaneously filters out waste products and supplies fresh components from an external "feeding" solution. Figure obtained by personal communication from Matsuda *et al.* (RIKEN Yokohama Institute).

2.2.3 Cloning of target DNA

The first step of cell-free systems is the preparation of target DNA suitable for use as a template for transcription. A common method begins with the amplification of target DNA by PCR (described in section 2.1.2) to produce discrete fragments which are then inserted into a cloning vector (see sections 2.1.5a – 2.1.5d) and sequenced (section 2.1.6) (steps 1 – 6, **figure 2.6**). An example of this procedure, and the one used to clone target DNA in section 4.2, is described in the following section.

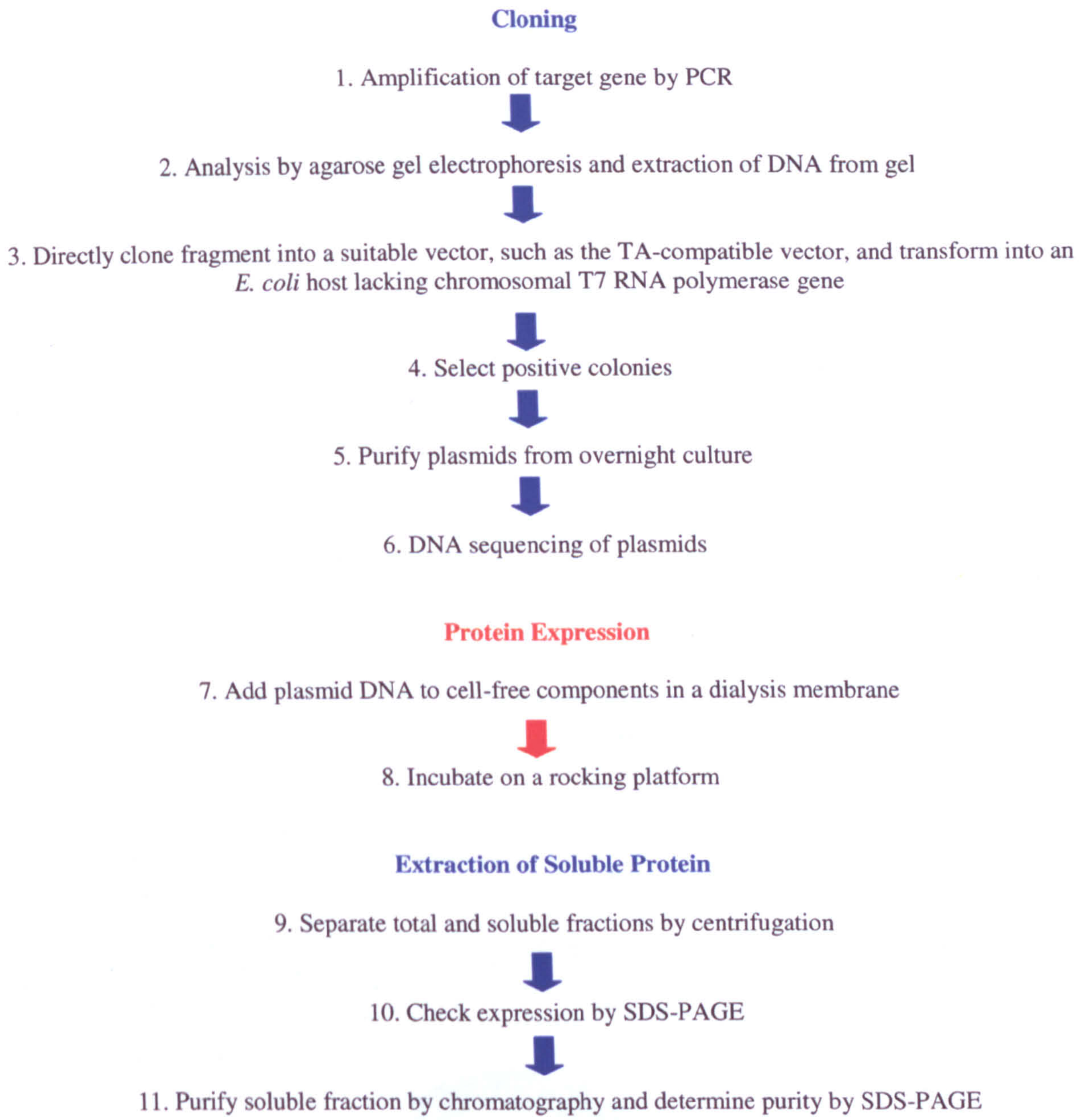

Cloning
1. Amplification of target gene by PCR
2. Analysis by agarose gel electrophoresis and extraction of DNA from gel
3. Directly clone fragment into a suitable vector, such as the TA-compatible vector, and transform into an *E. coli* host lacking chromosomal T7 RNA polymerase gene
4. Select positive colonies
5. Purify plasmids from overnight culture
6. DNA sequencing of plasmids
Protein Expression
7. Add plasmid DNA to cell-free components in a dialysis membrane
8. Incubate on a rocking platform
Extraction of Soluble Protein
9. Separate total and soluble fractions by centrifugation
10. Check expression by SDS-PAGE
11. Purify soluble fraction by chromatography and determine purity by SDS-PAGE

Figure 2.6: Schematic representation of a common procedure used for the expression of recombinant proteins using a bacterial (*E. coli*) *in vitro* system. This system was used to express proteins in section 4.2.

2.2.3a PCR

PCR is performed as described in section 2.1.2, however several additional sequences are included within the primers, resulting in a final PCR product which includes not only the target sequence, but also components essential for transcription and translation. Such components include a ribosome binding site, a T7 promoter and terminator, and stop codon sequences, if required. Tags to aid solubility or purification may also be engineered at this stage.

2.2.3b Invitrogen TOPO® TA cloning system

The resulting PCR fragment, purified by agarose gel electrophoresis, is then ligated into a suitable cloning vector. An example is the TOPO® TA Cloning System (**figure 2.7**) which exploits the dual restriction enzyme/ligase activity of toposiomerase I (*Vaccinia virus*) (Shuman, 1994). The enzyme cleaves a single strand of the plasmid at 5´-(C/T)CCTT-3´ and remains covalently bound to the phosphate group of the 3' thymidine. Re-ligation of the vector, and release of the enzyme, occurs upon addition of target DNA synthesised with single 3' adenosine overhangs.

Figure 2.7: The modified pCR®2.1-TOPO® cloning vector used as a template for the cell-free expression of proteins in section 4.2. Components essential for translation were introduced during PCR, see section 2.2.3a. An explanation of the additional components flanking the open reading frame (ORF) is also given in section 4.2.2a. Modified figure reproduced with permission from Invitrogen Ltd.

2.2.3c Cloning host

A host suitable for establishing recombinant plasmids in preparation for cell-free protein expression, is DH5α™ from Invitrogen. They support blue/white colony screening for the selection of positive transformants and have a high transformation efficiency, required for initial cloning. Insert stability is also improved by mutations in the *rec*A1 and *end*A1.

2.2.4 Optimisation of the cell-free reaction

The major benefit of cell-free expression systems is the ability to incorporate any number of additives, cofactors, or molecular chaperones, to test their effect upon total expression or solubility, on a very small scale (Betton, 2003 and Murthy *et al.*, 2004). Similar experiments in an *in vivo* system would be extremely time-consuming due to the larger volumes needed to obtain detectable quantities of target protein, together with the requirement for cell lysis. *In vitro*, this can easily be performed in a 96-well dialysis-plate format in a matter of hours. A common external reaction solution is provided to all of the wells, with each well containing an individual combination of additives and/or proteins. Detectable quantities of protein can be synthesised in an hour, when reactions are placed in a shaking incubator. Scaling up the reaction requires the use of individual dialysis cups or dialysis membranes, encased in a plastic box containing an external reaction solution.

Inclusion of non-ionic detergents, above their critical micelle concentration (CMC), in the cell-free reaction may help to solubilise proteins with transmembrane domains. These regions are incorporated into the hydrophobic core of the micelle, shielding the hydrophobic domains from solvent (Marston and Hartley, 1990). Detergents may also prevent aggregation and precipitation of proteins which do not contain transmembrane domains, by protecting hydrophobic domains from the solvent.

Molecular chaperones can be included in cell-free reactions to promote solubility and correct folding, by recognising hydrophobic residues or regions of unstructured backbone (Hartl and Hayer-Hartl, 2002). The groE system (groEL and groES) compartmentalises individual polypeptide chains, allowing them to fold correctly in isolation from one another, thus preventing aggregation (Wang and Boisvert, 2003), whilst the Hsp70 system (dnaK, the Hsp40 dnaJ, and the nucleotide exchange factor grpE, in *E. coli*) promotes folding through numerous cycles of ATP-dependent substrate binding and release (Hartl and Hayer-Hartl, 2002).

As with any expression system, the inclusion of certain additives such as detergents and molecular chaperones may adversely affect downstream applications, and so are only used when absolutely necessary. Problems may occur when attempting to separate additives from the solubilised protein, particularly when they shield hydrophobic regions from solvent. Some detergents, especially those with high CMC values, can often be removed by dialysis. Alternatively, detergents which form small micelles may be removed by gel filtration chromatography if the target protein's molecular weight is greater (Hjelmeland, 1990). Some detergents however, either cannot be completely removed from the system, or their removal causes target protein aggregation or precipitation.

Removal of molecular chaperones can sometimes be performed by chromatographic steps alone, however invariably this is unsuccessful. Incubating a protein-groE complex with ATP can induce a conformational change within the chaperones, which in turn may release the protein. Similarly ATP can liberate bound protein from Hsp70 complexes in two stages: firstly, ATP binds to dnaK and relieves dnaJ of the bound protein; and secondly, grpE catalyses the hydrolysis of ATP to ADP, inducing a conformational change in dnaK significant enough to release the protein (Hartl and Hayer-Hartl, 2002). As with detergents, removal of chaperones may result in the unfolding or aggregation of target protein.

The cell-free system can also be used to produce labelled proteins, such as selenomethionine-labelling in preparation for MAD-phasing, during crystallography (Kigawa *et al.*, 2002).

As with the *in vivo* system, it is necessary to optimise incubation parameters to obtain the maximum level of soluble protein. A typical large scale reaction requires 9 ml of internal and 90 ml of external solution and favourable conditions can produce milligram quantities of recombinant protein in several hours (steps 7 – 8, **figure 2.6**).

2.2.5 Extraction of protein

Soluble fractions can easily be obtained without the use of mechanical or chemical lysis due to the lack of whole cells within the reaction. Reaction solutions are transferred to a 96-well plate or a Falcon tube and centrifuged. The extent of expression is determined by SDS-PAGE (see section 2.1.9) and the soluble fraction can be purified by chromatographic methods. Purification tags such as histidine-tags are often incorporated during PCR, by the inclusion of such sequences within the oligonucleotide primers (section 2.1.2), and can provide a simple and generally highly effective first-stage purification (steps 9 – 11, **figure 2.6**).

2.3 Protein purification

2.3.1 Chromatography

Chromatography was first discovered in 1903 by Mikhail Tswett as a method for separating molecules through their specific interactions with porous solid matrices. Molecules solubilised in a mobile phase are passed through a column packed with a porous resin (the stationary phase). The properties of a molecule affect its interaction with the resin and subsequently determine its rate of migration through the column.

For simple separations, it is possible to perform chromatography manually on the bench, by using an air-filled syringe or a small pump to force the mobile phase through the matrix. This is particularly useful when purifying coloured proteins from a crude extract, as they may be identified visually.

Complex purifications are more conveniently performed using fast-protein liquid chromatography (FPLC), an automated system which precisely pumps samples at controlled flow rates through the matrix. Glass or plastic beads, 3 – 300 μm in diameter, coated with chromatographic media are packed into a column and attached to a system which incorporates an ultraviolet light source to measure the absorption spectra of eluted proteins. For purification of protein molecules, it is commonplace to measure absorbance at 280 nm, and peaks containing the protein of interest may be identified by SDS-PAGE.

Developments in FPLC technology have produced many commercially available, high specification systems which increase throughput and require less user intervention. The ÄKTA™ Explorer systems from Amersham Biosciences provides an automated platform for the purification of proteins through a series of chromatographic columns, from user-defined protocols.

2.3.1a Immobilised metal ion adsorption chromatography (IMAC)

Metal chelate affinity chromatography (IMAC) provides a simple first step purification of recombinant proteins engineered with an exposed polyhistidine tag at one end of the polypeptide chain. Ligands, such as Ni^{2+}, Cu^{2+}, or Zn^{2+} ions, which specifically bind to polyhistidine regions, are covalently bound to an inert matrix. As natural proteins do not generally bind with high affinity to these charged matrices, they are removed from the column in a low salt buffer. A competitive chelating reagent such as imidazole is added to the column to elute the non-covalently bound target protein. When polyhistidine tags are engineered well, the majority of contaminating proteins can be removed in just one step.

An example of commercially available metal chelate media, is the nickel-sepharose resin available from Amersham Biosciences. This consists of a chelating group, pre-charged with Ni^{2+} ions, coupled to highly cross-linked agarose beads.

2.3.1b Ion exchange chromatography

In contrast with metal affinity chromatography which can be performed with limited biochemical knowledge of the protein sample, it is generally necessary to know the isoelectric point of the target protein when performing ion exchange chromatography. Charged molecules bind to immobilised groups of opposite charge on a cellulose or agarose matrix. Proteins which are negatively charged below the pH of the buffer to be used, bind to cationic groups on an anion exchange column, and vice versa for positively charged cations. When the pI of the protein is not known, proteins may be separated by using a strong ion exchanger which functions over a wide pH range to determine the best system to use.

The column is washed with a low salt buffer and weakly bound proteins are removed from the matrix. The target protein is eluted in a gradient (linear or step-wise) of low to high salt buffer, which competitively binds to the charged resin, thereby releasing proteins at a rate dependent upon their binding affinity.

Examples of commercially available ion exchange resins include CM sepharose, a weak cation exchanger, and Q sepharose, a strong anion exchanger, both from Amsersham Biosciences. These resins are formed from 6 % highly cross-linked spherical agarose beads.

2.3.1c Gel filtration/size exclusion chromatography

By exploiting the porous nature of agarose beads, proteins can be separated according to their size and shape. The extent of cross-linking between agarose beads, and so pore size, is chosen dependent upon the desired range of molecular weights to be separated. When heterogeneous solutions are applied, smaller molecules pass through the pores and larger ones are excluded. The resulting effect being that larger molecules elute from the column at a faster rate than smaller molecules.

There is a linear relationship between the logarithm of the molecular mass of a protein and its relative elution volume from the column, hence it is possible to extrapolate the oligomeric state of a protein when the molecular weight is known. It is first necessary to calibrate the column by passing a heterogeneous solution of proteins of known molecular weight through the matrix. A calibration curve of a column is calculated by plotting Kav values (**equation 2.1**) of each known protein against their molecular weight.

Equation 2.1:

$$K_{av} = (V_e - V_o) / (V_t - V_o)$$

Where:

V_e = elution volume for the protein

V_o = void volume of the column (the volume of mobile phase between the stationary phase beads)

V_t = total bed volume of the column

2.4 DNA/protein characterisation

2.4.1 Bioinformatics

Bioinformatics brings together computer science, mathematics, and information theory, to enable the analysis of biological systems through the sharing of vast amounts of data. Such techniques have been essential in the collation of genomic data and have played significant roles in the progression of structural biology. A number of techniques used throughout this thesis are described briefly.

UniProt provides a comprehensive database of protein information, compiled from Swiss-Prot, TrEMBL, and PIR (http://www.ebi.uniprot.org), whilst the RCSB PDB (Protein Data Bank) provides structural information about biological macromolecules, highlighting their relationships to sequence, function, and disease (http://www.rcsb.org). PredictProtein enables the prediction of structure and function of entire proteins or particular domains, by comparing with similar sequences (www.predictprotein.org). The American NCBI (National Center for Biotechnology Information) database is a multi-purpose tool, providing access to information such as journal articles, protein/DNA sequences, and protein structures (http://www.ncbi.nlm.nih.gov). Global alignment of sequences can be performed using ClustalW, a multiple alignment tool which highlights similarities in sequences and also introduces gaps which represent evolutionary insertions or deletions (Thompson *et al.*, 1994 and www.ebi.ac.uk/clustalw/). Finally, SMART (Simple Modular Architecture Research Tool) can be used to estimate functional annotations of unknown protein sequences, based upon molecules of known structures (smart.embl-heidelberg.de/).

2.4.2 Electronic spectroscopy

Spectroscopic methods, and in particular electronic spectroscopy, are often used to characterise biological systems by studying the interaction of radiation with matter. In the ultra-violet/visible region, radiation may be partially absorbed by a molecule (chromophore), causing a rearrangement of electrons to a higher energy state (Hammes, 2005). This absorption is detected by measuring the difference in intensity between the light before and after it passes through the sample. Absorption can be quantified using the Beer-Lambert law, which states that there is a linear relationship between absorbance and concentration of an absorbing species (Lambert, 1760 and Beer, 1852). To determine absorption using the Beer-Lambert law, a known extinction coefficient is required. This is a constant value specific to the molecule of interest at a particular wavelength and is either experimentally derived or calculated using quantum mechanics (Hammes, 2005). Inaccuracies may occur when the light is not monochromatic or if the sample has aggregated.

The Beer-Lambert law, in terms of molarity, is written as:

Equation 2.2:

$$A = \varepsilon \times c \times l$$

Where:
A = experimentally derived absorbance
ε = wavelength dependent molar extinction coefficient in ($M^{-1}\ cm^{-1}$)
c = molar concentration of the protein
l = path length of the cuvette

Hence, the molar concentration of an unknown protein in solution may be calculated by:

Equation 2.3:

$$c = A / \varepsilon \times l$$

2.4.2a Protein quantification

Protein concentration may be estimated spectroscopically, by absorbance and colourmetric assays. Both methods can only estimate the concentration, particularly for impure samples. The concentration of a soluble protein may be calculated from its absorbance of ultraviolet light at 280 nm, as amino acids with aromatic rings absorb at this wavelength (Dunn, 1992). Chromophores which exhibit strong absorption in this region are phenylalanine, tyrosine, and tryptophan. The extent at which a specific pure protein, in a specific buffer, absorbs light at 280 nm can be calculated to yield a molar extinction coefficient (ε). It is therefore possible to calculate the concentration from the extinction coefficient specific to that protein (**equation 2.3**).

A second method for quantifying proteins in solution, and one which does not require knowledge of specific extinction coefficients, is the colourmetric assay developed by Bradford (Bradford, 1976). The Bradford assay measures the absorption change which occurs upon binding of protein to Coomassie brilliant blue G-250 dye. The red cationic form of the dye absorbs at a maximum of 465 nm, which shifts to the blue anionic form upon binding of certain amino acids, with an absorption maximum of 595 nm. The dye binds only to arginine, tryptophan, tyrosine, histidine, and phenylalanine residues.

Absorption measurements at 595 nm, for a series of standards including Bradford reagent, must first be performed. Bovine serum albumin (BSA) of known concentration is commonly used for this purpose, over a linear concentration range of 0.1 to 1.4 mg/ml. A standard curve is calculated by plotting absorbance at 595 nm against the known concentration. Unknown concentrations may then be calculated by performing the assay using several dilutions of the protein, complexed with Bradford reagent. Concentrations are extrapolated from the standard curve.

Protein concentration may also be determined, to a high degree of accuracy, using an amino acid analyser. This method determines the quantity of each amino acid within a protein, in four steps: 1. hydrolysis; 2. derivatization; 3. HPLC separation; and 4. data interpretation and analysis. Whilst this is the most accurate method of determining protein concentration, only a few laboratories are equipped with such a facility and is generally only used when precise quantification is required.

2.4.3 Circular dichroism

Circular dichroism (CD) exploits the differences in absorption of left and right handed polarised light by asymmetric or chiral molecules (Walker, 1998). Well ordered structures result in both positive and negative signals, whilst irregular structures give a zero signal. Chromophores within the protein, namely the peptide bonds, absorb in the "far" UV region (170 to 250 nm), the resulting data of which can be used to predict secondary structure. This is because α-helices, β-sheets, and random coils, give rise to a characteristic spectrum which can be interpreted using various algorithms to estimate the percentage of each structural element. This method cannot however be used to identify specific residues involved in such formations. CD performed at synchrotron radiation sources provides significantly improved signal to noise ratios than bench sources and so enables a greater accuracy of predictions (Jones and Clarke, 2004), particularly at shorter wavelengths.

A further use of CD occurs in the "near" UV region (250 to 350 nm), whereby aromatic residues and disulphide bonds contribute to the spectra. Signals in this region can indicate the correct folding of protein and can be used to assess the effects of buffer, PH, salt, and ligand-binding, amongst other variables (http://www.ap-lab.com/circular_dichroism.htm).

2.4.4 Electron paramagnetic resonance

Electron paramagnetic resonance (EPR) spectroscopy studies the effect of radiation on molecules within a strong magnetic field (Hames, 2005). An EPR signal arises due to an unpaired *d*-orbital electron within a molecule, which in the case of haem proteins is supplied by the ferric iron (Fe^{3+}), which has one unpaired electron in its outermost shell. Such molecules are paramagnetic and give rise to an EPR signal (Hammes, 2005), whilst molecules which have a full complement of electrons in the outer shell are "EPR-silent".

Chapter 3 – Theoretical and experimental background to protein crystallography

3.1 Introduction

The structural study of biologically important molecules can be achieved using many different methods with varying levels of resolution. Techniques include electron microscopy, nuclear magnetic resonance (NMR), small-angle X-ray scattering (SAXS), and protein crystallography (PX), the most powerful of which, PX, produces high resolution three dimensional models of the molecule of interest. Highly pure protein is generally required to grow crystals suitable for PX and the production of such crystals is responsible for one of the major bottle-necks of the technique. Another consideration is that crystal packing may also affect the true structure of the protein. Both NMR and SAXS yield structural information of proteins in solution, thereby removing the requirement for crystallised sample material. NMR allows small molecules to be visualised in motion within a solvent, however sensitivity may be compromised.

Unlike the universal helical structure of DNA (Watson and Crick, 1953), every protein has a unique configuration which may be only partially conserved within molecules of homologous sequence or function. The first protein structures, myoglobin and haemoglobin, were not published until 1960 and further growth within the field remained slow until the surge in computing power in the 1970s and the availability of Synchrotron radiation in the 1980s (Giacovazzo *et al.*, 2002). Currently there are 38, 620 protein structures deposited in the Protein Data Bank (PDB) database (Berman *et al.*, 2003), 46 % of which were deposited in the last five years (2000 to 2005). Of the structures available, 84.8 % were determined by single-crystal X-ray crystallography, 14.6 % by NMR, and just 0.3 % by electron microscopy.

3.2 X-ray diffraction

Protein crystallography centres around the principal that ordered structures such as crystals, which contain a regular lattice of molecules, scatter bombarded X-rays from atomic electrons. The diffracting rays can either constructively or destructively interfere with each other, producing an interference pattern. In order for multiple X-rays to be scattered in phase (constructively), they must satisfy Bragg's law (**equation 3.1**) (Bragg, 1912). This is shown in **figure 3.1**, which illustrates diffraction from multiple planes (of atoms) within a crystal, separated by a distance (*d*). The radiation will travel different distances dependent upon the distance between the diffracting planes. For the waves to scatter in phase, these distances must be integral (*n*) multiples of the wavelength (Hammes, 2005).

Equation 3.1:

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

Where *n* is an integer number, λ is the wavelength of the X-rays, *d* is the spacing between planes within the crystal lattice, and θ is the angle between the incident ray and the scattering planes.

Figure 3.1: Diffraction of X-rays from a crystal lattice. The parallel lines represent planes of atoms within a crystal, which are separated by a distance (*d*), and the angle at which radiation interacts with the crystal is shown as θ. Figure taken from Hammes, 2005.

The direction and intensity of the scattered X-rays are measured by an automatic detector (such as a CCD or image plate) and computational methods are used to convert this diffraction image into a three dimensional electron density map. Briefly, once the program has found a value for all reflections in the reciprocal indices (h,k,l), this can be considered as the equivalent real-space direction along the crystal's axes (x,y,z) when a Fourier transform is applied to the structure factors, F(hkl). Structure factors include three components, the frequency (pre-determined by the source wavelength), the amplitude (calculated from the measured intensity I_{hkl}), and the phase (determined by phasing methods, see section 3.5). The structure factors may be expressed as:

Equation 3.2:

$$F(hkl) = F_{hkl} \exp(i\alpha_{hkl})$$

Where F_{hkl} is the amplitude and α_{hkl} is the phase of the structure factor for each reflection. F_{hkl} is directly measured from the experimental data as it is related to the intensity of the reflections (I_{hkl}), however the phase information (α_{hkl}) must be computationally derived.

The structure factors, obtained from a crystallographic experiment, may be represented by:

Equation 3.3:

$$\mathbf{F}(hkl) = \sum_{j=1}^{N} \mathbf{f}_j \exp(2\pi i(hx_j + ky_j + lz_j))$$

Where h,k,l define the coordinates in reciprocal space of a particular reflection (the Miller indices), x_j, y_j, z_j are the coordinates for the j^{th} atom, and f_j is the scattering factor for the j^{th} atom.

A Fourier transformation of the structure factors, F(hkl), may be expressed as:

Equation 3.4:

$$\rho(xyz) = \frac{1}{V} \sum_h \sum_k \sum_l |F(hkl)| \exp[-2\pi i(hx + ky + lz) + i\alpha(hkl)]$$

Where $p(xyz)$ is the calculated electron density map, the $F(hkl)$ amplitudes are the sum of all f_{hkl} values for individual atoms in a unit cell, and V is the volume of the unit cell.

The detailed description of the theory and principles of X-ray diffraction have been well documented and are beyond the scope of this work, so will not be discussed further (Blundell and Johnson, 1976, Drenth, 1999, Ladd and Plamer, 1994, and Stout and Jensen, 1989).

3.3 X-ray diffraction data collection

3.3.1 Crystal growth

The production of suitable sample material and the growth of single crystals accounts for one of the major bottle-necks in protein crystallography. Highly pure protein (> 95 %) is generally required for the growth of crystals suitable for X-ray diffraction experiments, however in some cases contamination may actually improve crystallisation. Sample purity is usually estimated by separating contaminating proteins from the target by electrophoresis and comparing the subsequent band intensities.

A number of modern techniques exist for the crystallisation of macromolecules and are described further in the literature (Ducruix and Giege, 1992 and McPherson, 1999). The commonly used method of vapour diffusion gradually removes water from a drop containing both protein and a precipitant, at a start concentration just below that required to precipitate the protein. This occurs by placing the drop over a reservoir containing precipitant solution in a closed system, allowing for equilibration. The drop can be suspended over the reservoir on a siliconised glass coverslip, as in hanging-drop vapour diffusion, or placed on a small plastic bridge over the reservoir, as in sitting-drop systems (**figure 3.2**). Further techniques include sandwich drop and microdialysis crystallisation. In sandwich drop crystallisation, protein is combined with precipitant and placed between two siliconised coverslips, with a small gap at each end to enable diffusion. Microdialysis crystallisation involves the gradual exchange of two precipitant solutions of varying ionic strength/pH. This technique can be used for proteins which require high concentrations of salt for solubility (**figure 3.2**).

The system slowly reaches supersaturation, whereby evaporation from the drop sufficiently increases precipitant concentration, so that crystal formation can occur. If this process occurs too quickly, such as if the start precipitant concentration is too high, the protein will precipitate out of solution and no crystalline growth will be observed (**figure 3.3**). Alternatively when the process is successful, clusters of protein form nuclei from which crystals may grow. When multiple nuclei are present within a drop, many small microcrystals may be formed, which can then be individually used as seeds in fresh drops to induce the formation of larger crystals. This method can be used to produce single crystals and to improve the overall quality.

Conditions used to obtain initial crystals can be optimised by varying factors such as pH, salt concentration, and temperature. Small molecular additives may also be included, which can manipulate sample-sample/sample-solvent interactions, as well as the water structure (Hampton Research).

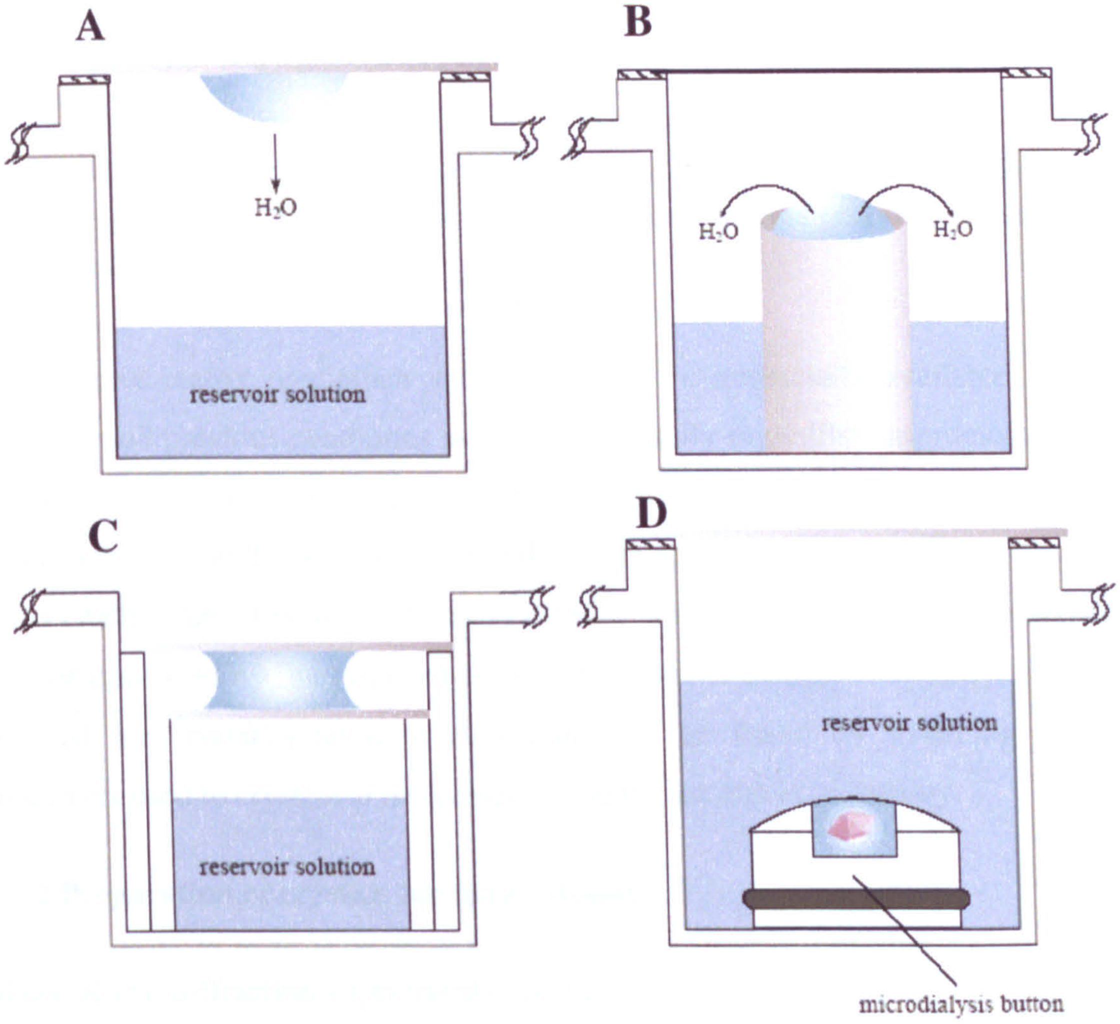


Figure 3.2: Diagrammatic representation of four common protein crystallisation techniques. **(A)** Hanging-drop vapour-diffusion, **(B)** sitting-drop vapour-diffusion, **(C)** sandwich-drop crystallisation, and **(D)** microdialysis crystallisation. Figures reproduced with permission from Hampton Research Corporation.

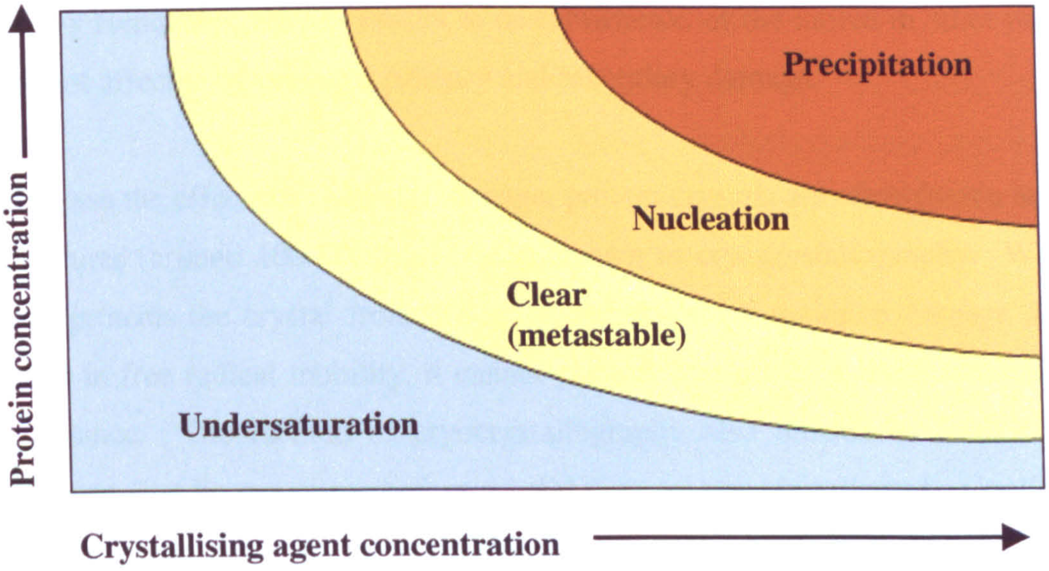


Figure 3.3: Idealised phase diagram showing the probability of nucleation in relation to supersaturation of the crystallisation system. The blue region represents undersaturation and yellow/brown regions represent supersaturation.

Many sparse-matrix precipitant screens are now commercially available, formulated based around previous conditions used to successfully crystallise macromolecules (such as Hampton Research, Molecular Dimensions, and Qiagen Nextal). These are available in multiple formats for both manual and robotic screening and provide a starting point for crystallisation of novel proteins. Potential "hits" obtained from these screens can then be optimised by finer manual screening. Another method to identify the condition required for crystallographic growth can often be found by screening around the conditions used to crystallise molecules of significant sequence identity.

3.3.2 Preparation of crystals for X-ray studies

Many X-ray diffraction experiments are successfully performed at room temperature, however this procedure does little to minimise the damage caused by high-intensity radiation, particularly when using macromolecular crystals. Such crystals may only survive for several minutes in a synchrotron beam, usually not enough time to collect sufficient data. Primary radiation damage occurs when X-rays cleave bonds within the crystal, producing free radicals (Gonzalez and Nave, 1994). These highly reactive molecules can diffuse through the crystal's solvent channels, causing secondary damage by reacting with other molecules, destabilising the crystal and disrupting the precise crystal lattice (Nave, 1995). Finally tertiary damage, also described as the "domino effect" by Henderson (1990), results in destabilisation of the lattice in other parts of the crystal not affected by previous primary and secondary damage.

To decrease the effects of radiation damage, protein crystals are often frozen at very low temperatures (around 100 K), in a process known as cryocrystallography. Whilst such cooling protects the crystal from the latter two stages of radiation damage due to the decrease in free radical mobility, it cannot prevent formation of these molecules in the first instance. The method of cryocrystallography also introduces further problems which must first be overcome before a valid data set can be collected. Cooling to low temperatures can result in the formation of crystalline ice which disrupts the crystal lattice. Soaking the crystals in mother liquor containing a cryoprotectant (such as glycerol, ethylene glycol, PEG, or MPD), before flash-freezing in a stream of liquid nitrogen, can help prevent this.

Cryoprotectants, when applied correctly to the crystal without causing damage, form an amorphous phase both within and around the crystal upon flash-freezing, thus protecting the precise internal lattice. In such situations, cryoprotectants also reduce background scattering from water and provide effective platforms for the storage and transport of crystals. However when the conditions are not exact, flash-freezing can affect the order of the crystal, resulting in a higher mosaicity. Mosaicity refers to the angular measurement of order within a crystal lattice. In some cases mosaicity, and even resolution, can be improved by reannealing the crystal, whereby the cryostream is interrupted briefly before remeasuring the data (Samygina *et al.*, 2000 and Ellis *et al.*, 2002). Detwinning of crystals has also been resolved using this method.

3.3.3 X-ray radiation sources

3.3.3a Conventional sources

For the purpose of crystallography, X-ray radiation can be obtained from either conventional laboratory or synchrotron sources. Laboratory generators are both weaker in intensity and allow less possible experimental wavelengths than synchrotron sources, however they provide a cheaper and convenient in-house alternative. Such laboratory sources can be further categorised into sealed tube and rotating anode generators, the latter of which produces significantly more intensity than sealed tube sources

(Giacovazzo *et al.*, 2002). Sealed-tube generators consist of a high-voltage power supply (40 to 50 kV) which accelerates cathode-generated electrons through a vacuum, towards a fixed metal anode plate, producing X-rays which escape from the tube through perpendicular beryllium windows. The transformation efficiency of electrons into X-rays is just 0.1 %, the limiting factor being the efficiency of the system used to cool the anode.

As the name suggests rotating anode sources overcome this by the continual movement of the metal anode target, thereby allowing greater power to be applied, resulting in a higher intensity X-ray output. Whilst copper is the most commonly used metal target, other materials such as molybdenum and tungsten can be used, resulting in a limited number of wavelength variations.

3.3.3b Synchrotron sources

Synchrotron radiation sources (**figure 3.4**) produce a wide-ranging spectrum of light and produce X-rays which are at least 10^5 times more intense than those from rotating anode generators, allowing for the collection of very high resolution data (Giacovazzo *et al.*, 2002). Such facilities were first developed in the 1960s and have since been improved to include sources of different "generations" (Helliwell, 1992). Initial first generation sources produced synchrotron radiation merely as a by-product of high-energy particle physics. The first dedicated synchrotron facility, the SRS at Daresbury in Cheshire, paved the way for other second generation sources whereby X-rays were principally generated via bending magnets. Further modifications have since been added to these facilities to improve intensity, such as the addition of undulators, wigglers, and wavelength shifters. Finally, third-generation sources such as the ESRF in Grenoble operate with significantly greater flux and brilliance, further increasing the potential quality of data obtainable.

When the direction of a charged particle beam changes, the electrons or positrons are accelerated, emitting a continuous spectrum of electromagnetic radiation, at a wavelength characteristic of the bending magnet. Particles are first accelerated in a linear accelerator and then in a booster ring accelerator, before injecting into a storage ring. Bending magnets within this large polygonal chamber ensure the particles circulate continuously (Giacovazzo *et al.*, 2002). All synchrotron chambers are kept under the best possible vacuum. Electrical current within the storage ring decreases over time, caused by interactions between the accelerated particles and contaminating atoms, due to this deficiency. Radio-frequency transmitters provide the energy source for Synchrotron radiation sources.

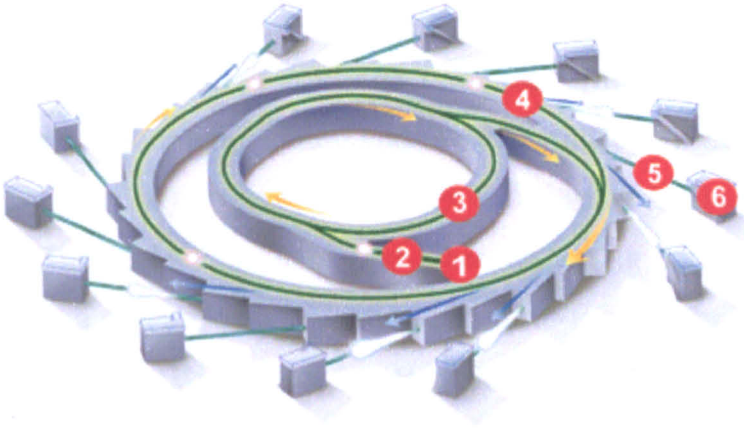


Figure 3.4: Generalised diagram showing the layout of a synchrotron radiation facility. Electrons are generated from a high-voltage source **(1)** and are accelerated by the linear accelerator (linac) **(2)** and then the booster ring **(3)**, before injecting into the storage ring **(4)**. An individual beamline is shown at position **5** and a user station at position **6**. Figure taken from the Canadian Virtual Science Fair, 2005 (www.virtualsciencefair.org/2005/shar5a0/how_does_a_synchrotron_work.htm).

Bending magnets alter the direction of the beam by accelerating particles towards the centre of the ring, thus emitting radiation tangentially, which is then focused by quadrupolar magnets (Giacovazzo *et al.*, 2002). Crystal monochromators are used to reduce the electromagnetic spectrum to a user-defined wavelength, thereby supporting many different experimental applications. Monochromators consist of one or two stable crystals, commonly silicon, orientated with one face parallel to a major set of crystal planes (Giacovazzo *et al.*, 2002). Such crystals diffract the incoming beam at a wavelength determined by the angle of the crystals scattering plate, as described by Bragg's law (**equation 3.1**). Multiple crystal monochromators, as implemented on station 10.1 at the SRS, result in very narrow bandwidths which prevent movement of the X-ray beam during wavelength modifications and allow for an increased flux and rapid tunability around an absorption edge (Cianci *et al.*, 2005).

Further modifications to synchrotron sources include the addition of wigglers and undulators, a number of magnets with alternating polarities, which are positioned within the straight sections of the storage ring (Giacovazzo *et al.*, 2002). These alter the characteristics of the radiation by shortening the wavelength and increasing acceleration, thereby increasing intensity.

3.3.4 Data collection

Protein crystals can be mounted ready for X-ray exposure in one of two ways, dependent upon the temperature at which data are to be collected. Micro-capillary mounting is used when collecting at room temperature and involves 'injection' of the crystal into a small capillary. The capillary is sealed with wax and then secured onto a goniometer head using putty. The second method, used when collecting data at 100K, involves soaking the crystal in a cryoprotectant solution and flash-freezing in liquid nitrogen, as described in section 3.3.2. The crystal is suspended in a loop and then immersed in cryoprotectant. The loop is mounted onto a magnetic base and secured to a goniometer head in a nitrogen cryostream.

Once the crystal has been mounted correctly, a preliminary assessment to determine the unit cell parameters and the approximate quality and resolution of diffraction, is performed. The crystal, attached to a goniometer, is placed in the pathway of the X-ray beam and several test images are recorded and processed. If these parameters are satisfactory, I_{hkl} intensities of the recorded reflections are measured to as high a resolution as possible. If they are not, additional crystals must be selected and analysed. Preliminary analysis can also allow appropriate experimental parameters such as total oscillation angle, crystal to detector distance, exposure time, and the wavelength at which data are to be determined. Data are then reduced and individual reflections are indexed, resulting in known unit cell parameters and space group. The data are then scaled and merged, before determining the phase of each reflection (see sections 3.4 - 3.5).

The data used to determine the structure described in chapter 5 were collected on the MAD station 10.1 at the SRS, Daresbury (Cianci *et al.*, 2005) using a CCD detector.

3.4 Data processing

Once experimental data have been collected, a number of processing steps must first be performed before structure determination can occur. Many computational programmes are available to perform these tasks, including Mosflm (Leslie, 1992) and the program which was used to process all data in this work, HKL2000 (Otwinowski and Minor, 1997).

3.4.1 Data reduction

HKL2000 performs the first stage of processing, data reduction, in two steps. Firstly, all possible indices for all measured reflections are identified, to determine correct h, k, l values. An estimation of the unit cell dimensions, crystal orientation, and symmetry point group can be made from this auto-indexing step. Further parameters such as mosaicity, spot shape, and beam position, are then refined to optimise the fit of the predicted diffraction pattern against the observed experimental pattern. Statistical methods are used to assess the quality of this fit, one weighted factor is defined as χ^2 (in both directions of the two-dimensional plane). Smaller values indicate a better agreement between the two data sets and values below 2.0 are generally considered to be acceptable.

Secondly, diffraction spot intensities are accurately recorded using a profile-fitting integration method in HKL2000. The spot profiles over a specified area of the detector are averaged and each spot is assigned a profile based upon this value. Signal to noise ratio is estimated by measuring a small area of diffraction background. Each reflection is assigned as either fully recorded, where the spot is entirely measured in one image, or partially recorded, where the spot is measured over a number of images. When partially recorded reflections are observed, the full profile of the spot must be calculated from the sum of each these images. A wider oscillation range of data collection may be performed to ensure the spot is collected in its entirety, however this can result in an increase in background and may also result in the overlap of different reflections. The optimum oscillation range can be identified during preliminary analysis of the crystal.

3.4.2 Data scaling and merging

The next stage of data processing involves scaling the data from each individual image and merging the data into a single dataset. The SCALEPACK program within the HKL2000 package (Otwinowski, 1993) was used during these stages throughout this thesis. Inconsistencies may arise during data collection, such as the varying of X-ray beam intensity, detector sensitivity, or differing thickness and imperfections visible within the crystal during rotation. Scaling increases the consistency of the dataset by merging identical reflections of the same index (h,k,l) from different images and assigning them identical intensities. Identical reflections may also be merged from different diffraction patterns obtained from multiple crystals, or from single crystals collected at more than one resolution. The factor by which datasets are scaled (the scale factor) is calculated as described by Fox and Holmes (1968), whereby the scale between a reference image and the last image (I) is given by:

Equation 3.5:

$$G_j = K_i \exp\left(\frac{-2B_i \sin^2\theta}{\lambda^2}\right)$$

Where K_i and B_i are the scale and temperature factors between the images.

These factors are then applied to the data and a single value is assigned to each h, k, l reflection. The merged intensities are then converted to structure factors and the magnitude is determined by the French and Wilson method using the CCP4 program TRUNCATE (French and Wilson, 1978 and CCP4, 1994). Finally, an overall temperature factor may also be approximated (Wilson, 1949).

3.5 Phasing of macromolecular diffraction data

As described in section 3.2, a three dimensional electron density map of the crystal can be obtained by performing a Fourier transform of the structure factors, F(hkl). The magnitude term of these factors can be calculated directly from the diffraction pattern, however the phase term is lost, and so must be determined by alternative means. A number of methods exist to overcome this “phase problem”, the most popular of which are discussed here.

3.5.1 Molecular replacement (MR)

Molecular replacement seeks to determine the phases of an unknown crystal structure by comparison with a known homologous structure, as first described by Rossmann and Blow (1962). If a significant level of structural identity exists between the two molecules, estimates of the unknown’s phases can be sufficiently accurate to allow structure determination. The “phasing model”, or the known structure, is placed into the crystal system of the unknown structure (the “experimental” model) and the correct positioning of this allows for the estimation of phase factors. Computational methods are used to orient the phasing model using six transformation parameters, three rotational and three translational, which are split into two processes for computational efficiency. The cross-rotation function first defines the relative orientations of the experimental data within the phasing model and the translational function then attempts to position the correctly orientated model into a unit cell of the experimental crystal.

The Patterson function drives MR and is used to represent the summation of the product of electron densities in a crystal at points separated by a vector (u,v,w). The output of this function, a Patterson map (**equations 3.6 – 3.7**), is a three dimensional plot of the function with axes (u,v,w). Vectors between atoms within the crystal are represented as vectors between an origin and peaks on the map, with peak height proportional to the square of the atomic number.

Equation 3.6:

$$P(u,v,w) = \int_{v} \rho(x,y,z)\rho(x+u,y+v,z+w)dv$$

Which is expressed for crystallographic purposes as:

Equation 3.7:

$$P(u,v,w) = \frac{1}{V}\sum\sum\sum |F|^2 \cos 2\pi(hu+kv+lw)$$

The maximum radius for which electron density is considered for inclusion in a Patterson map can be chosen so as to optimise the probability of finding a correct orientation. As Patterson vectors correspond to distances between atoms, those within a 2r radius around the Patterson map are generally regarded as intramolecular, however a small number of close intermolecular vectors may also be included in this radius. The value, r, corresponds to the maximum atomic displacement from the centre of mass.

3.5.1a The cross-rotation function

The first stage of MR, cross-rotation, involves the superimposition of the phasing Patterson map over that of the experimental one, in an attempt to determine the relative orientations of the two models. The correlations of the two maps are calculated at each set of angles as the phasing model is rotated three dimensionally through a specified radius (r) around the origin. This rotational transformation between the two maps represents transformations between the two structures and ideally only includes intramolecular vectors. The optimum orientation is generally calculated in reciprocal space using the Crowther-Blow algorithm (**equation 3.8**) (Crowther and Blow, 1967), however a more accurate orientation can be obtained using the slower real-space method.

Equation 3.8:

$$R = \int_u P_2(X_2)P_1(X_1)dx_1$$

Where $P_1(X_1)$ is the Patterson map calculated from the diffraction data, $P_2(X_2)$ is the Paterson of the rotated phasing model, and the integral is carried out over a volume of *u*.

3.5.1b The translation function

The second stage of MR phasing, translation, then attempts to position the correctly orientated model into a unit cell of the experimental crystal using only intermolecular (cross) Patterson vectors. The crystal's space group determines the position of specific Harker sections, around which Patterson vectors are clustered. The optimised orientation of the two Paterson maps, obtained from cross-rotation, is placed at its origin within the unit cell. Cross Patterson vectors are measured at each position and screened against the experimental Patterson data for similar vectors:

Equation 3.9:

$$T(t) = \int P(u)P(u,t)du$$

Finally, the position and orientation of the phasing model is refined to improve its fit to the experimentally derived structure factors. This stage is performed automatically at the end of molecular replacement when using the MOLREP program (Vagin and Teplyakov, 1997). Relative phases can then be calculated from the positioning of the phasing map within the unknown's unit cell and these, together with the observed (magnitude) structure factors, enable calculation of an electron density map.

3.5.2 Isomorphous replacement

Isomorphous replacement, the first method developed to solve the phase problem in protein structures, involves the use of two crystals, one soaked in a heavy metal solution (such as platinum, mercury, or uranium) and another "native" crystal (Green *et al.*, 1954). Both crystals are required to be highly similar (isomorphous) with regards to unit cell parameters and symmetry. Such metals bind tightly to one or more sites within the asymmetric unit and form a major constituent of the overall X-ray scattering. The differential scattering (a "difference" Patterson map) between the two crystals forms the basis of this technique. The positions of the incorporated metal atoms can be inferred from this Paterson map, allowing for subsequent approximation of phase for each reflection using the Harker construction (Harker, 1956). The major downside of this technique is the potential for disruption of the structure by metal incorporation and the difficulty in obtaining isomorphous crystals.

Whilst it is possible to solve the phase problem using just one heavy metal derivative, as in the case of Single Isomorphous Replacement (SIR), it is more common to use two or more derivatives, known as Multiple Isomorphous Replacement (MIR). The use of multiple metals increases the accuracy of phase determination and hence increases the chance of successful structure determination.

3.5.3 Anomalous scattering

Anomalous scattering is another phasing method which relies upon the properties of metals. Advances in synchrotron technology have enabled the development of beamlines with tuneable X-ray wavelengths, which are exploited in anomalous scattering experiments. All metals absorb X-rays at a specific wavelength, known as their absorption edge, and anomalous dispersion occurs when the wavelength is close to this value. Multiple Anomalous Dispersion (MAD) phasing uses two or more wavelengths for data collection and is particularly useful for phasing metalloproteins, which may intrinsically contain sufficient metal content so as not to require additional soaking. The most commonly used procedure of heavy atom incorporation for anomalous phasing is the substitution of methionine with selenomethionine during protein expression. Production of selenomethionine-labelled proteins is well documented, both in *in vivo* (Leahy *et al.*, 1992) and *in vitro* (Kigawa *et al.*, 2002) systems. Finally, the availability of intense beam at wavelengths > 2 Å has enabled the use of native sulphur phasing, without the need for additional metals (Dauter *et al.*, 1999).

3.6 Structure refinement and validation

Whichever phasing method is used, positioning of the model is then refined to improve the fit of the calculated structure factors with the experimental data. With the exception of structures determined by the molecular replacement method, it is first necessary to build a model into the electron density. This can be performed using graphical interfaces such as O (Jones *et al.*, 1991) or Coot (Emsley *et al.*, 2004). The resulting model, or the model produced from molecular replacement, is subjected to cycles of automatic refinement by programs such as REFMAC5 using the CCP4 suite (Murshudov *et al.*, 1997), which refines the model and re-calculates the electron density map, before being re-built manually. Refinement is generally performed using stereochemical restraints, due to the low ratio of observed (Fo) to refinable parameters (x, y, z, B). The extent of these restraints depends upon the resolution at which data are collected, with lower resolution data sets requiring higher weightings. A crystallographic R-factor is used to monitor the agreement between the model and the experimental data during these cycles of refinement. The value of R is inversely proportional to this agreement and is expressed as:

Equation 3.10:

$$R = \frac{\sum_{hkl} \left| |Fobs| - k|Fcalc| \right|}{\sum_{hkl} |Fobs|}$$

Where *R* is the R-factor, *k* is the scaling factor, *Fobs* is the observed experimental structure factors, and *Fcalc* is the calculated structure factors.

A second value, R-free, allows for the cross-validation of refinement to prevent over-fitting of the data (Brünger, 1992). Such over-refinement of the data is particularly apparent when working at low resolution, due to the reduced number of observed reflections. Prior to any refinement of the data, a small random subset of reflections (about 5 %) are removed from the refinement process and used to calculate the R-free value as described below:

Equation 3.11:

$$R_T^{free} = \frac{\sum_{hkl \subset T} \left| |Fobs| - k|Fcalc| \right|}{\sum_{hkl \subset T} |Fobs|}$$

The R-free value is expected to decrease during a successful progression of refinement. A typical over-refined data set will yield a low R-factor with an unchanged or increased R-free value. However in all refinements, the R-free value generally remains approximately 2 to 5 % higher than the R-factor.

Throughout the refinement process a number of other validation methods are performed to monitor the quality of the model. The programs PROCHECK (Laskowski *et al.*, 1993) and WHATIF (Vriend, 1990) assess the overall structure and individual residues using a number of stereochemical tests. Root-mean-square (rms) deviations of the model's bond lengths and angles from accepted values are measured throughout refinement and provide an additional validation parameter (McRee and David, 1999). A library of standard bond characteristics, derived from simple organic compounds (Engh and Huber, 1991), provide the reference values for each bond. Rms values for bond lengths and angles of < 0.02 Å and < 3 ° respectively are considered to be well refined, however these vary dependent upon the resolution at which data are collected.

Ramachandran plots are also used to assess the stereochemical validity of the structure by plotting the phi φ and psi ψ dihedral angles of each residue (Ramachandran *et al.*, 1963). Only certain combinations are sterically feasible and values should fall within the accepted range as defined by the plot. As glycine residues lack side chain atoms, a wider range of angle combinations are allowed. A further parameter, the estimated standard uncertainty (ESU) value, estimates the overall coordinate error. This value estimates the data-only contribution to the positional uncertainty for an atom with a temperature factor equal to the Wilson B value for the whole molecule (Cruickshank, 1996).

3.6.1 REFMAC5 refinement

All refinement in this thesis was performed using REFMAC5, which employs a maximum likelihood strategy for refinement, whereby the best model is that which is most statistically consistent with the experimental data. The detailed working of the REFMAC5 program is described elsewhere (Murshudov *et al.*, 1997) and so will only briefly be described here. When refinement results in a change to the model which is more probable, the likelihood increases, providing a measurable quantity of improvement. Errors from both the model and the measurements are also taken into account when calculating probability. Such errors in the model decrease with each successful refinement cycle, which results in a sharpening of the probability values and a subsequent increase in likelihood, up to a maximum value.

The program, ARP/wARP (Lamzin, 1993) can be used within REFMAC5, to enable the addition of solvent molecules during the refinement process. Further development of the REFMAC5 program has enabled the use to TLS refinement (Winn *et al.*, 2001) and user-defined weight restraining, which takes into account any new data added during the refinement process.

Chapter 4 – Production of proteins from *Mycobacterium tuberculosis*

4.1 Introduction and target selection

A number of protein targets from *M. tb* were selected for trials using an *E. coli*-based cell-free expression system (section 4.2), with the intent of producing soluble proteins suitable for downstream applications such as protein crystallography. The expression of several targets was also attempted using an *in vivo E. coli* system (section 4.3), to enable comparisons to be made between the two systems.

Individual target proteins were initially selected based on the research interests of members of the NWSGC. A total of 28 targets were chosen for high-throughput trials using a cell-free expression system. Whilst a small percentage of the targets were hypothetical proteins, the majority were assigned putative functions based upon sequence analysis carried out by Cole *et al.* (1998 and 2002) and subsequent homology database searches using BLAST (Altschul *et al.,* 1990). See **table 4.1** for recent functional annotations of the 28 targets.

A review of current literature, and in particular a paper by Sassetti and Rubin in 2003 (see section 1.1.3a), enabled the selection of additional targets, unaffiliated with the NWSGC. To ensure existing crystal structures were not available for these targets, each was screened against the Tuberculosis Structural Genomics Consortium (TBSGC) database, the Protein Data Bank (Berman *et al.*, 2003), and the National Center for Biotechnology Information (NCBI) database. Prediction of membrane domains using the SOSUI Secondary Structure database (Hirokawa *et al.*, 1998) also eliminated targets from the selection process. Finally, due to our group's interest in metalloproteins, targets were chosen based upon their predicted metal content (see **table 4.2** for functional annotations). This resulted in the selection of eight further metalloprotein targets.

Rv[1]	Gene[2]	Functional Annotation[2]
0153c	0153c	Phosphotyrosine protein phosphatase
0171	mce1C	MCE-family protein
0185	0185	Zinc metallopeptidase
0247c	0247c	Succinate dehydrogenase
0359	0359	Zinc metallopeptidase
0505c	serB1	Phosphoserine phosphatase
1388	mihF	Integration host factor
1407	1407	FMU protein
1942c	1942c	Conserved hypothetical protein
1967	mce3B	MCE family protein
2060	2060	Conserved integral membrane protein
2229c	2229c	Conserved hypothetical protein
2234	ptp	Phosphotyrosine phosphatase
2305	2305	Hypothetical conserved protein
2547	copG	Transcriptional regulator protein
2711	IdeR	Iron-dependent repressor protein
2718c	2718c	Conserved hypothetical protein
2776c	2776c	Oxidoreductase
2981c	2981c	D-alanine-D-alanine ligase
2986c	hupB	DNA binding protein
3042c	serB2	Phosphoserine phosphatase
3070	3070	Conserved integral membrane protein
3628	ppa	Inorganic pyrophosphatase
3712	3712	Ligase
3717	3717	Hypothetical protein
3836	3836	Zinc metalloprotease
3867	3867	Conserved hypothetical protein
3915	3915	Hydrolase

Table 4.1: The 28 *Mycobacterium tuberculosis* targets chosen for high-throughput trials using a cell-free expression system. [1]Gene number assigned by sequence analysis (Cole *et al.*, 1998). [2]Current gene and functional annotations taken from the Tuberculosis Structural Genomics Consortium (TBSGC) database.

Rv[1]	Gene[2]	Function Annotation[2]	Metal[3]	Progress (from TBSGC)[4]
0670	end	Endonuclease IV	Zinc	Cloned 2002
0950c	0950c	Probable metalloprotease	Unknown	Targeted 2003
1589	bioB	Biotin synthetase	Iron-sulphur	Cloned 2002
2388c	hemN	Oxygen-independent coproporphyrinogen III oxidase	Iron-sulphur	Expressed 2001
2845c	proS	Prolyl-tRNA synthetase	Zinc	Targeted 2004
3534c	3534c	4-Hydroxy-2-oxovalerate aldolase	Unknown	Not targeted
3545c	3545c	Cytochrome P450 125	Iron (haem)	Not targeted
3781	3781	ATP-binding protein ABC transporter	Unknown	Not targeted

Table 4.2: The eight *Mycobacterium tuberculosis* targets, found to be essential for *in vivo* infection, chosen for expression trials using a cell-free system. [1]Gene number assigned by sequence analysis (Cole *et al.*, 1998), [2]Current gene and functional annotations taken from the TBSGC database. [3]Metal requirement predicted by literature review of homologoues. [4]Current progress of the target by members of the TBSGC.

4.2 Cell-free protein expression

4.2.1 Introduction

Two visits were made to the Protein Research Group at RIKEN, to exploit the cell-free technique over a period of eighteen weeks. During the first six week visit, a high-throughput approach was employed to screen 28 NWSGC targets from *Mycobacterium tuberculosis* (*M. tb*) (**table 4.1**). The constructs of all targets were prepared by PCR and ligated into a cloning vector. After selection of positive clones, targets were screened for expression on a small scale. Those targets which yielded soluble protein were progressed into large scale synthesis to yield milligram quantities of protein, which were subsequently purified using affinity chromatography. This work is described in sections 4.2.2 to 4.2.3.

A different strategy was employed during the second twelve-week visit. The eight new targets (**table 4.2**), together with five NWSGC targets which were insoluble during the first-round of cell-free trials (Rv0185, Rv0247c, Rv2776c, Rv3717, and Rv3915), were selected. This reduction in the number of targets allowed a more thorough approach to the optimisation of expression conditions. To examine the effect of additives on total yield and solubility, metals, molecular chaperones, and detergents were added to the cell-free reaction solutions. A number of soluble targets were expressed on a large-scale, together with a further four targets which were produced on a large-scale during the first visit (Rv2229c, Rv2547, Rv2981c, and Rv3836), to replenish protein stocks for crystallisation trials. This work is described in sections 4.2.4 to 4.2.5.

4.2.2 High-throughput cell-free expression of 28 *Mycobacterium tuberculosis* targets: Methods

All PCR primers described in section 4.2.2a were designed by Takashi Yabuki and Yukiko Fujikura. Cloning steps described in section 4.2.2b were performed by Takayoshi Matsuda and positive clones (section 4.2.2c) were identified by Eiko Seki, Masaomi Ikari, and Fumiko Hiroyasu. Optimisation of unsuccessful PCR reactions using DMSO (section 4.2.4a), was performed by Dr. John Hall from De Montfort University, whilst at the RIKEN Yokohama Institute. The *E. coli* S30 cell-free extracts were prepared by Natsuko Matsuda and Natsumi Suzuki (Kigawa *et al.*, 2004).

The recipes for buffers and media described in this section are given in appendix 2.

4.2.2a PCR

M. tb genes were amplified directly from genomic DNA using a 2-step PCR method (**figure 4.1**) (Yabuki *et al.*, personal communication). Initial target-unique primers, also encoding a linker sequence, were used to amplify the genes from genomic *M. tb* DNA (H37Rv). 300 ng of genomic DNA (obtained from Colorado State University) was included in a typical 20 µl PCR reaction solution of: 50 nM of each primer (Invitrogen); 0.2 mM of each dNTP (dATP, dCTP, dGTP, and dUTP); 0.5 U expand HiFi *Taq* DNA polymerase (Roche); and 1 x HiFi buffer (Roche). PCR cycling parameters are shown in **table 4.3**.

The resulting PCR product (PCR 1) was used as a template for the second PCR step (PCR 2). This was performed using a 'universal' primer, which annealed to additional fragments

(T7T and T7P) included in the second PCR reaction solution, via universal linker sequences. The 5' T7P fragment encoded a T7 promoter, a ribosome binding site, a native histidine tag (HAT^{TM} tag, BD Biosciences Clontech), and a *Tev* protease cleavage site. The T7T fragment encoded two 3' downstream stop codons and a T7 terminator. These fragments annealed not only to the universal primers, but also to the linker sequence on the PCR 1 product, thereby synthesising the gene of interest linked to essential transcription and translation components, and providing the starting template for cell-free expression. The PCR 1 product was diluted five-fold and included in a reaction solution of 1 μM 'universal' primer, 0.2 mM of each dNTP, 0.5 U expand HiFi *Taq* DNA polymerase, and 1 x HiFi buffer (both Roche), together with 50 pM of the T7T and T7P fragments.

PCR reactions were performed using a PTC-200 thermocycler (MJ Research). PCR products were analysed by agarose gel electrophoresis on a 1 % agarose gel containing 0.5 μg/ml ethidium bromide (Sigma) and 1 x TBE buffer. Gels were run at an electrical potential of 150 V and viewed under ultraviolet fluorescence.

5 % DMSO was included during PCR for targets which were not successfully amplified. Templates with a high GC content usually require a higher strand separation temperature, leading to a decrease in product yield. Chemicals such as DMSO are employed to disrupt base pairing, thereby reducing the temperature requirement.

4.2.2b Cloning of target DNA

Although it was possible to synthesise target proteins directly from PCR products, polymerases can introduce single point mutations. To ensure no mutations had occurred, PCR products were inserted into a cloning vector in preparation for DNA sequencing. 1 μl of the final PCR product (PCR 2) was combined with 0.5 μl pCR$^{®}$2.1-TOPO$^{®}$ plasmid (Invitrogen), 0.5 μl salt solution (Invitrogen), and 1 μl Milli-Q water on ice. Reactions were incubated at room temperature for fifteen minutes. 1 μl of the ligation reaction was added to 15 μl of DH5α competent cells (Invitrogen) and incubated on ice for five minutes. Plasmids were used to transform the cells by heat-shock at 42 °C for 45 seconds before returning on ice for two minutes. After the addition of 150 μl SOC medium, cells were incubated at 37 °C for 40 minutes.

The culture was spread onto an LB-Kan-IPTG-Xgal plate and incubated overnight at 37 °C. Blue-white screening was used to identify positive clones. For each target, 12 single white clones were used to independently inoculate 1 ml of super broth including 5.6 mM glucose and 25 µg/ml kanamycin in a 96 deep-well plate. The plate was incubated overnight at 37 °C in a shaker incubator. Plasmid DNA was purified using the standard Wizard® Miniprep kit protocol (Promega) and then sequenced to identify positive clones (see section 4.2.2c).

Positive clones were grown in 200 ml of LB-Kan overnight and purified using a Wizard® Midiprep kit (Promega) on a vacuum manifold, following the manufacturer's protocol. DNA was eluted with 300 µl MilliQ water, pre-heated to 70 °C. Purified plasmids were visualised by agarose gel electrophoresis and yield calculated by Picogreen analysis (Invitrogen), before storing at – 20 °C.

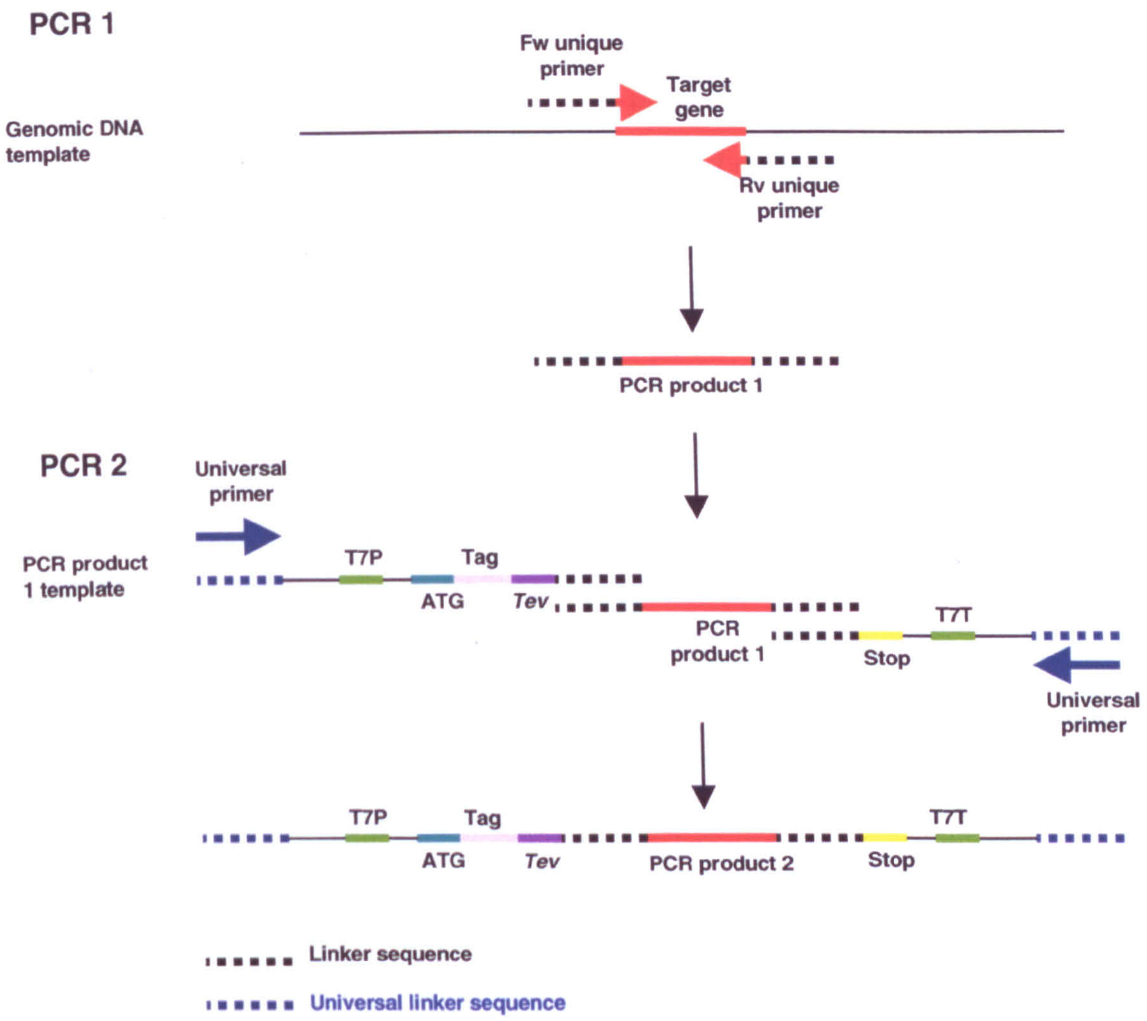


Figure 4.1: The 2-step PCR procedure used to amplify target genes from *Mycobacterium tuberculosis* DNA, in preparation for cell-free expression.

		Cycling parameters		
PCR number	**Template**	**Temp. (°C)**	**Time (s)**	**Cycles**
1	Genomic DNA	94	2 min	1
		94 60 72	30 30 60	20
		94 60 72	30 30 60 +5 sec/cycle	20
		72	7 min	1

		Cycling parameters		
PCR number	**Template**	**Temp. (°C)**	**Time (s)**	**Cycles**
2	PCR 1 product	94	2 min	1
		94 60 72	30 30 90	10
		94 64 72	30 30 90 +5 sec/cycle	20
		72	7 min	1

Table 4.3: PCR cycling parameters used to amplify target DNA in preparation for cell-free expression.

4.2.2c DNA sequencing

Sequencing reactions were carried out in two 96-well PCR plates, denoted forward and reverse. 0.2 µM of sequence primer (M13) was mixed on ice with 2 µl plasmid DNA, 1 x sequencing buffer (Applied Biosystems), 1 µl BigDye^{TM} v3.1 terminator premix (Applied Biosystems), and 5.34 µl Milli-Q water. The sequencing reaction was performed using hot-start PCR at an initial denaturation temperature of 96 °C for 30 seconds, followed by 25 cycles of: 1 °C per second to 96 °C; 96 °C for 10 seconds; 1 °C per second to 50 °C; 50 °C for 5 seconds; 1 °C per second to 60 °C; and 60 °C for 4 minutes.

Completed sequencing reactions were placed on ice and 2.5 µl of 125 mM EDTA pH 8.0, followed by 25 µl of 100 % ethanol were added. Reactions were incubated at room temperature for 15 minutes before centrifuging at 5, 650 rpm for 30 minutes at 4 °C. Pellets were washed with 30 µl of 70 % ethanol and centrifuged at 5, 650 rpm for 10 minutes. Pellets were resuspended in 20 µl of HiDi formamide (Applied Biosystems) and boiled for 2 minutes at 95 °C. PCR products were sequenced using an in-house ABI PRISM® 3100 Genetic Analyzer (Applied Biosystems).

Sequence data was aligned with the correct nucleotide sequence (TBSGC) for each target using ClustalW (Thompson *et al.*, 1994) on MacVector software (Accelrys). Clones with nonsense, missense, and read-through mutations were discarded. Positive clones (mutation-free or in some cases, those with silent mutations) were progressed into expression trials.

4.2.2d Cell-free protein expression: Initial screening

Initial screening for expression of target proteins was carried out using a 96-well micro-dialysis plate (15 kDa molecular weight cut off) from PCR 2 templates. Determination of successfully amplified PCR products (correct size) was assessed by agarose gel electrophoresis. Template DNA was not sequenced at this stage as the cloning steps detailed in section 4.2.2b were conducted in parallel to initial screening experiments.

A 30 µl internal reaction solution contained 9 µl *E. coli* S30 extract (Kigawa *et al.*, 2004); 1 µl of PCR 2 product; 1.8 mM DTT; 1.2 mM ATP; 0.8 mM each of CTP, GTP, and UTP;

0.64 mM 3', 5'-cyclic AMP; 68 μM L(-)-5-formyl-5,6,7,8-tetrahydrofolic acid; 66.6 μg/ml T7 RNA polymerase; 175 μg/ml *E. coli* total tRNA (Roche); 1.5 mM of each amino acid; 80 mM creatine phosphate (Roche); 0.25 mg/ml creatine kinase (Roche); 9.28 mM magnesium acetate; 27.5 mM ammonium acetate; 200 mM potassium glutamate; 0.05 % sodium azide; 58 mM Hepes-KOH pH 7.5; and 4.0 % PEG 8000 (Sigma).

A 130 ml universal external solution contained the same components as the internal solution with the exception of PCR 2 product, T7 RNA polymerase, creatine kinase, and tRNA. The *E. coli* S30 extract was substituted with S30 buffer, 14 mM magnesium acetate, and 1 mM DTT. The cell-free dialysis reactions were incubated at 37 °C for 8 hours, whilst the external solution was mixed on a magnetic platform.

After the incubation, 30 μl of buffer CF-A was added to the reaction solutions and soluble fractions were obtained by centrifugation at 5, 650 rpm for 5 minutes at 4 °C, after removal of 5 μl for the total fraction. 5 μl was removed for the soluble fraction and 25 μl of MilliQ water was added to both fractions. Due to the inclusion of PEG in the reaction solution, fractions were acetone precipitated for 5 minutes on ice and centrifuged for 30 minutes, to ensure good electrophoresis resolution. Pellets were dried at 65 °C for 20 minutes and resuspended in 30 μl of 1 x SDS sample buffer. Fractions (10 μl) were analysed by sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE), as described by Laemmli (1970). Molecular weight markers from Sigma (SDS7) allowed for the estimation of protein molecular weight.

4.2.2e Cell-free protein expression: Optimisation

Optimisation of cell-free reaction conditions were performed using the dialysis method from PCR 2 templates. Capped dialysis cups with 15 kDa cut-off membranes were placed inside 1.5 ml screw-top tubes containing magnetic stirrers. Dialysis cups contained the internal reaction solution of 30 μl and a 300 μl external reaction solution was housed in the screw-top tube. Reactions were stirred on an incubated magnetic platform.

Metal compounds were incorporated into the cell-free reactions for targets which gave no or insoluble expression during initial screening. 50 μM of $ZnSO_4$ and $FeSO_4$ were added to both internal and external reaction solutions for the zinc and iron metalloproteins, respectively. Both metals were added separately to the proteins with unknown metal affinity. Metal solutions were also included in the reaction solutions of soluble targets thought to bind zinc or iron, to identify any effect on yield. Reactions were set up using PCR 2 templates and incubated overnight at 37 °C.

A time-course study was set up to examine the effect of incubation time on expression for a number of targets that were synthesised incorrectly under standard reaction conditions. Rv0359, Rv2547, and Rv2718 were synthesised from PCR 2 templates at 37 °C and solubility was assessed at varying time points between 0 to 24 hours.

To enable the possibility of MAD-phasing during crystallography, it was necessary to substitute methionine with selenomethionoine within the cell-free reaction. As insolubility of target protein can occur as a result of selenomethionine oxidation, methionine was substituted with 1.5 mM selenomethionine to assess the effect on protein solubility before progressing to large-scale synthesis. Soluble proteins were synthesised from plasmid templates (1 µg/ml of pCR®2.1-TOPO®-target was included in the internal reaction solution) at 30 °C for 8 hours.

4.2.2f *Tev* cleavage study

All target proteins were constructed to include a His-tag to aid purification. An upstream *Tev* protease cleavage site was also included to enable cleavage of the His-tag (~ 15 kDa) following affinity purification. To achieve optimal cleavage, the concentration of *Tev* protease and incubation parameters were analysed using Rv3628 as a test target. The cell-free expression system was scaled up to a 3 ml internal reaction solution, dialysed against a 30 ml external solution overnight at 30 °C, with gentle shaking. Rv3628 was synthesised from plasmid template.

1.6 ml of TALON™ Superflow cobalt resin (BD Biosciences Clontech) was added to the soluble fraction, obtained by centrifuging at 8, 000 rpm for 5 minutes at 4 °C. The slurry was incubated at room temperature for 20 minutes, to allow the protein to bind to the resin. The slurry was then loaded onto a TurboFilter plate (Qiagen), pre-equilibrated with buffer CF-A. A vacuum was applied to the plate at a rate of 1 to 2 drops per second (flow-through fraction). The resin was washed three times with 3.2 ml of the same buffer (wash fraction) and His-tagged proteins were eluted in 4 ml of CF-B. Elution fractions were concentrated to 2 ml using Amicon 5 kDa cut-off centrifugation filters at 4, 000 x g and split into 150 µl aliquots. Aliquots were incubated with varying concentrations of *Tev* protease at 4 °C and 30 °C for 3 hours to overnight (**table 4.4**). The extent of His-tag cleavage was analysed by SDS-PAGE.

Time	**Temperature (°C)**	**Concentration Of *Tev* protease (µg/ml)**
3 hours	4 30	5 10 20
	4 30	5 10 20
	4 30	5 10 20
Overnight	4 30	5 10 20
	4 30	5 10 20
	4 30	5 10 20

Table 4.4: Optimisation of *Tev* cleavage conditions for Rv3628. Aliquots of Rv3628, synthesised by cell-free expression, were incubated with *Tev* protease and the extent of His-tag cleavage was analysed by SDS-PAGE.

4.2.2g Large scale cell-free expression and purification

Nine targets, which were successfully expressed in small-scale trials, were synthesised on a large scale. These were Rv2229c, Rv2234, Rv2547, Rv2711, Rv2981c, Rv3042c, Rv3628, Rv3836, and Rv3867. All targets, with the exception of Rv3836 which was synthesised in the presence of 50 µM $ZnSO_4$, were expressed without the inclusion of additives in the reaction solutions.

Cell-free reactions were scaled up to include 18 ml (2 x 9 ml) internal and 180 ml (2 x 90 ml) external reaction solutions in 15 kDa dialysis membranes. Methionine was substituted with 1.5 mM selenomethionine (Sigma) to allow for MAD phasing. Proteins were synthesised at 30 °C for 8 hours, with gentle shaking.

The synthesised His-tagged proteins were initially purified using vacuum-manifold cobalt-affinity chromatography as described in section 4.2.2f, with the following exceptions: 4.8 ml of cobalt resin was added to each of the soluble fractions and proteins were washed with 9.6 ml of buffer CF-A; His-tagged proteins were eluted in 12 ml of buffer CF-B and concentrated to 6 ml. After a three hour incubation of eluate at 4 °C with 20 μg/ml *Tev* protease, proteins were desalted (500 mM to 0.5 mM imidazole) by buffer exchange into buffer CF-A, using Amicon 5 kDa concentrators. His-tags were then separated from target protein by affinity chromatography.

Further purification steps to obtain protein of adequate purity for crystallisation trials were carried out using an ÄKTA Explorer system (Amersham Biosciences) at CCLRC Daresbury Laboratory. Proteins were buffer exchanged into buffer CF-C, concentrated to 10 ml, and loaded onto a 5 ml anion exchange column (HiTrap™ Q Sepharose™ HP IEX, Amersham Biosciences) equilibrated in buffer CF-C. Target proteins were eluted in a gradient of 0 to 100 % buffer CF-D. Proteins were concentrated to between 0.5 and 2 ml and then loaded onto a Superdex 75 10/300 column (Amersham Biosciences), equilibrated in buffer CF-E. The gel-filtration step was not necessary for targets Rv2547, Rv2981c, and Rv3836, Rv3867, which were judged to be > 95 % pure by SDS-PAGE, following anion-exchange chromatography. These proteins were buffer exchanged into buffer CF-E. All buffers were passed through a 0.2 μm Whatman filter membrane and degassed prior to use. Following analysis by SDS-PAGE, fractions of the required purity were pooled and protein quantification was performed using a standard Bradford assay (Bradford, 1976). 100 μl of protein sample was incubated with 5 ml of a five-fold diluted Bradford assay reagent (BioRad) for 30 minutes at room temperature. Absorbance readings at 595 nm were obtained using plastic disposable cuvettes. Protein concentration was calculated by performing a typical bovine serum albumin (BSA, Sigma) standard curve. Proteins were concentrated to 5 mg/ml and stored at 4 °C.

4.2.3 High-throughput cell-free expression of 28 *Mycobacterium tuberculosis* targets: Results

In order to present these results in a concise manner it was thought appropriate to only include complete photographic evidence for targets which were expressed in milligram quantities, due to the large-scale of the study. Further gel photographs are included in appendix 1 and a summary of results is provided in **table 4.5**.

				Optimisation result	
Target	**Successful PCR (+ DMSO)**[1]	**Positive clone obtained**[2]	**Initial screening result**[3]	**Metals**[4] **($FeSO_4$ and $ZnSO_4$)**	**Timecourse**[5]
Rv0153c	Yes (+)	No	No expression	No expression	
Rv0171	No (+)	No	No expression	n/a	
Rv0185	Yes	Yes	Insoluble	Insoluble ($ZnSO_4$)	
Rv0247c	Yes	Yes	Insoluble	Insoluble ($FeSO_4$)	
Rv0359	Yes (+)	Yes	No expression	No expression	Smaller MW unknown protein present at all time-points except 0 hour
Rv0505c	Yes (+)	Yes	No expression	No expression ($ZnSO_4$)	
Rv1388	Yes	Yes	Soluble	n/a	
Rv1407	Yes (+)	Yes	Insoluble	Insoluble	
Rv1942c	No (+)	No	No expression	n/a	
Rv1967	No (+)	No	No expression	n/a	
Rv2060	Yes	Yes	No expression	No expression	
Rv2229c	Yes	Yes	Soluble	n/a	
Rv2234	Yes (+)	Yes	Soluble	n/a	
Rv2305	Yes (+)	No	No expression	No expression	
Rv2547	Yes	Yes	Soluble	n/a	Smaller MW unknown protein present at all time-points except 0 hour

Rv2711	Yes	Yes	Soluble	Soluble ($FeSO_4$)	
Rv2718c	Yes	Yes	Soluble	n/a	Larger MW unknown protein present from 2 hour time-point onwards
Rv2776c	Yes	Yes	Insoluble	Insoluble ($FeSO_4$)	
Rv2986c	Yes	Yes	No expression	No expression	
Rv3042c	Yes	Yes	Soluble	n/a	
Rv3070	Yes	Yes	Insoluble	Insoluble	
Rv3628	Yes	Yes	Soluble	n/a	
Rv3836	Yes	Yes	Soluble	Soluble ($ZnSO_4$)	
Rv3867	Yes	Yes	Soluble	n/a	
Rv2981c	Yes (+)	Yes	Soluble	n/a	
Rv3712	Yes (+)	Yes	No expression	n/a	
Rv3717	Yes	Yes	Insoluble	n/a	
Rv3915	Yes	Yes	Insoluble	n/a	

Table 4.5: Results summary for the cloning and small-scale expression of 28 *Mycobacterium tuberculosis* targets using the cell-free system. [1]Single PCR products of correct size, synthesised in the presence or absence of DMSO (section 4.2.3a). [2]Clones free from nonsense, missense, and read-through mutations (section 4.2.3a). [3]Expression result from initial screening using standard conditions (section 4.2.3b). [4]Expression result from optimisation of reaction conditions including metal compounds (section 4.2.3c). [5]Expression result of time-course study for targets where non-target or non-*E. coli* proteins were expressed (section 4.2.3c).

4.2.3a Cloning of target DNA

Single products of correct size were amplified by PCR for seventeen of the targets. An example of successfully amplified PCR products is shown in **figure A1**, appendix 1. Due to the characteristically high GC content of the *M. tb* genome, it was necessary to add DMSO to the remaining targets during PCR (**table 4.5**). This procedure was successful for all but three of the target genes, Rv0171, Rv1942c, and Rv1967. Positive PCR products were used as templates for small-scale cell-free expression (sections 4.2.3b to 4.2.3c). Positive clones were obtained for all targets except Rv0153c, Rv0171, Rv1942c, Rv1967 and Rv2305. Positive clones were used as cell-free expression templates for those targets produced on a large-scale (section 4.2.3d).

4.2.3b Cell-free protein expression: Initial screening

Initial expression trials performed using standard conditions from PCR 2 templates resulted in the synthesis of soluble protein for nine targets, insoluble protein for six, and no product for the remaining targets. For those targets which were not amplified correctly by PCR, no protein was produced due to errors in the template used for expression (the PCR product). Inclusion of DMSO during PCR for these targets produced discrete products and resulted in the soluble expression of two further proteins and the insoluble expression of one. Disregarding the three targets for which positive PCR 2 templates were not available, only seven targets gave no expression. Results are shown in **figure 4.2** and are summarised in **table 4.5**.

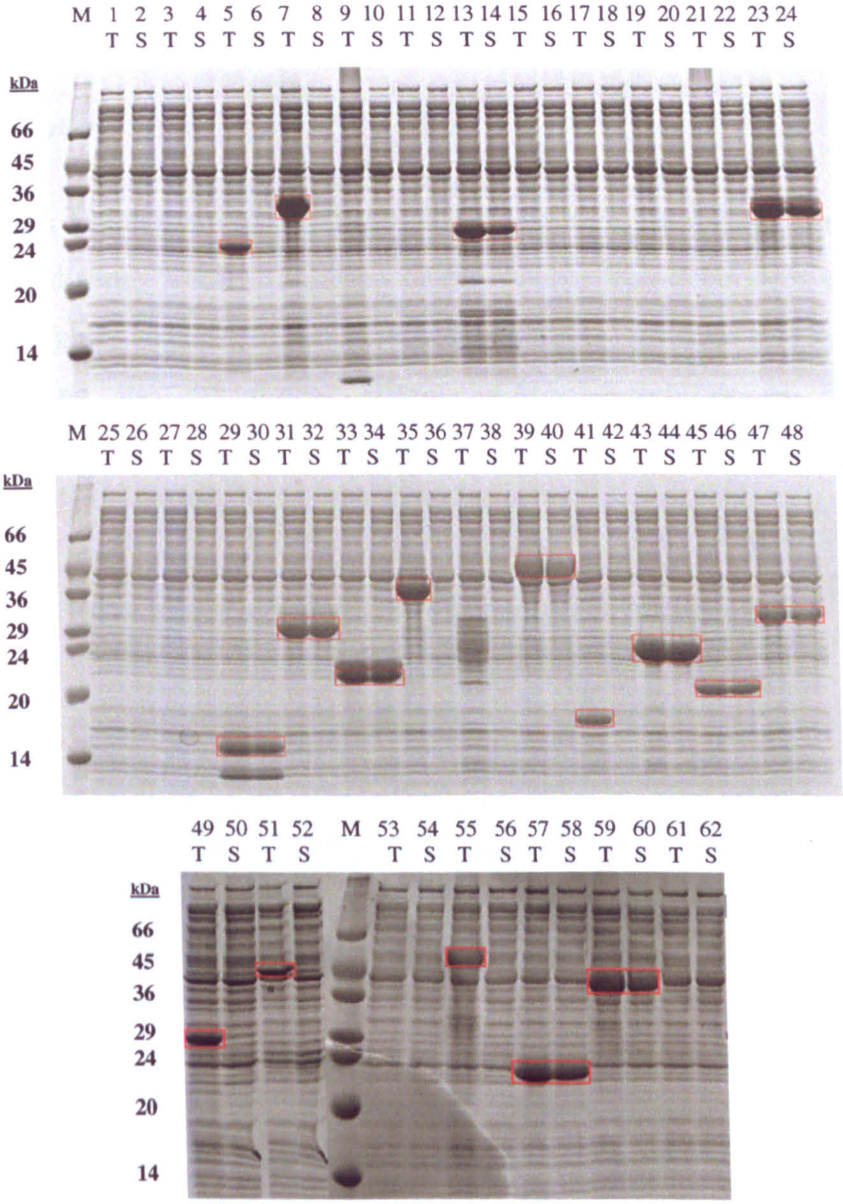

M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
T S T S T S T S T S T S T S T S T S T S T S T S
kDa
66
45
36
29
24
20
14
M 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
T S T S T S T S T S T S T S T S T S T S T S T S
kDa
66
45
36
29
24
20
14
49 50 51 52 M 53 54 55 56 57 58 59 60 61 62
T S T S T S T S T S T S T S
kDa
66
45
36
29
24
20
14

Figure 4.2: 10 % SDS-PAGE of initial cell-free expression screening from PCR 2 templates (section 4.2.2d). Total extracts (T) and soluble fractions (S) from reactions incubated at 37 °C for 8 hours. **Rv0153c:** in lanes 1 - 2; **Rv0171:** 3 - 4; **Rv0185:** 5 - 6; **Rv0247c:** 7 - 8; **Rv0359:** 9 – 10; **Rv0505c:** 11 - 12; **Rv1388:** 13 - 14; **Rv1407:** 15 - 16; **Rv1942c:** 17 - 18; **Rv1967:** 19 - 20; **Rv2060:** 21 - 22; **Rv2229c:** 23 - 24; **Rv2234:** 25 - 26; **Rv2305:** 27 - 28; **Rv2547:** 29 – 30; **Rv2711:** 31 – 32; **Rv2718:** 33 - 34; **Rv2776c:** 35 - 36; **Rv2986c:** 37 - 38; **Rv3042c:** 39 - 40; **Rv3070:** 41 - 42; **Rv3628:** 43 - 44; **Rv3836:** 45 - 46; **Rv3867:** 47 - 48; **Rv3717:** 49 - 50; **Rv3915:** 51 - 52.

Initial cell-free expression screening from PCR templates synthesised with DMSO. **Rv0505c:** 53 - 54; **Rv1407:** 55 – 56; **Rv2234:** 57 – 58; **Rv2981c:** 59 – 60; **Rv3712:** 61 - 62. Molecular weight marker in lanes M. Target proteins highlighted in red.

4.2.3c Cell-free protein expression: Optimisation

Metal incorporation

For those targets which were insoluble or gave no protein, no change was observed upon addition of metal ions. Total expression of the soluble iron-dependent repressor protein, Rv2711, was slightly reduced in the presence of $FeSO_4$, however this may be due to an inaccuracy during the experiment (**figure 4.3**). Solubility of Rv3836, a zinc metalloprotease, remained unchanged upon addition of $ZnSO_4$ (**figure 4.3**).

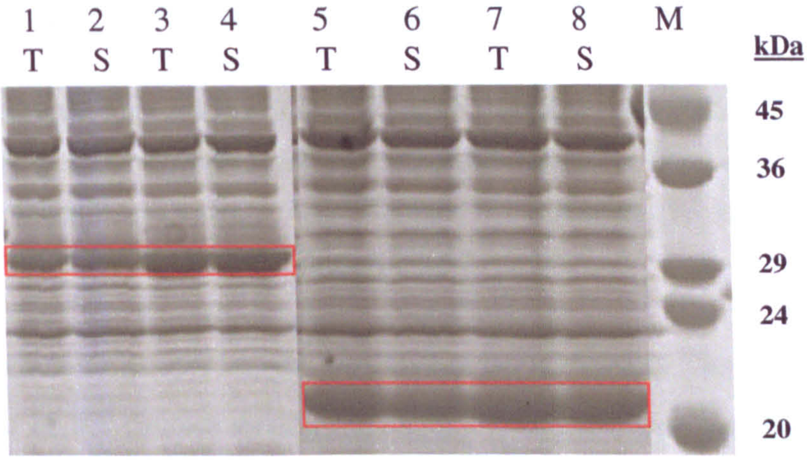


Figure 4.3: 10 % SDS-PAGE from optimisation of cell-free expression conditions by the addition of metal compounds (50 μM) (section 4.2.2e). Total extracts (T) and soluble fractions (S) from small-scale reactions incubated at 37 °C overnight. **Rv2711** with: **$FeSO_4$**, lanes 1 - 2; **No metal**, 3 - 4. **Rv3836** with: **$ZnSO_4$**, 5 - 6; **No metal**, 7 - 8. Molecular weight marker in lane M. Target proteins highlighted in red.

Time-course expression study

For a number of targets, non-*E. coli* proteins were expressed which did not correspond to the expected molecular weights of the target proteins. In these instances, a time-course expression study was performed to determine whether these proteins were degradation products of the proteins of interest.

Although single PCR products were obtained for both Rv2547 PCR reactions, a lower molecular weight protein was over-expressed in addition to the 9.5 kDa target protein. This band was detectable by SDS-PAGE from the 2 hour time-point and was most noticeable after the overnight incubation, coinciding with the dramatic decrease in the amount of full sized target protein at this time-point. This leads to the assumption that the lower molecular weight band was a product of target protein degradation, rather than one of co-expression (**figure 4.4**).

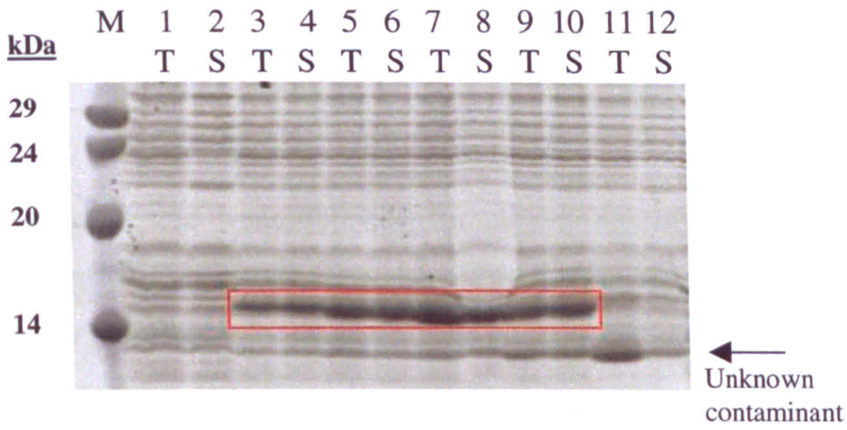


Figure 4.4: 10 % SDS-PAGE from cell-free time-course expression study of **Rv2547** (section 4.2.2e). Total extracts (T) and soluble fractions (S) from small-scale reactions incubated at 37 °C over 0 to 24 hours. **0 hours**: lanes 1 - 2; **2 hours**: 3 - 4; **4 hours**: 5 - 6; **6 hours**: 7 - 8; **8 hours**: 9 - 10; **24 hours**: 11 - 12. Molecular weight markers in lane M. Rv2547 highlighted in red.

No band of the correct molecular weight was visible by SDS-PAGE for Rv0359, possibly due to degradation of the protein during synthesis or to incomplete amplification of the entire gene by PCR. However, a non-*E. coli* band was observed at approximately 10 kDa. It was hypothesised that the band was a target protein degradation product and the possibility that reducing reaction incubation time would eliminate this was investigated. This band was not present at the 0 hour time-point, emphasising its non-*E. coli* origin, however it was present at every other time-point measured (**figure A2**, appendix 1). No band of correct molecular weight was visible, possibly due to degradation of the target protein during synthesis. Another explanation is that multiple bands were obtained by PCR for Rv0359, suggesting that the gene was not amplified correctly. Repeated cell-free expression from a single-band PCR product (PCR 2), synthesised with DMSO, did not result in production of either target protein or the 10 kDa unknown protein.

A larger molecular weight, non-*E. coli* protein, was co-expressed with Rv2718. This band was present after a two hour incubation and increased in intensity up to the overnight time-point. As the DNA template, obtained by PCR, was a discrete band and the protein appears to be larger than the expected molecular weight of the target protein (so is not degradation), it remains unclear why this contaminant is present (**figure A2,** appendix 1).

Effect of selenomethionine substitution

No change in overall expression or solubility was noticeable after substitution of methionine with selenomethionine.

Tev cleavage study

Rv3628 bands of reduced size were obtained for all conditions tested, signifying that cleavage of the His-tags had occurred. The cleaved tags were not visible by SDS-PAGE after 3 hours, however cleavage was evidenced by the presence of two differently sized Rv3628 bands. This absence of cleaved His-tags may be due to a sample loading error during SDS-PAGE. Incubating target proteins with 20 µg/ml *Tev* protease for 3 hours was deemed to provide sufficient cleavage, and so this method was chosen to conserve experimental time. As no difference was seen between 4 °C and 30 °C incubations, the lower temperature was selected to avoid potential degradation or aggregation of target protein (**figure 4.5**).

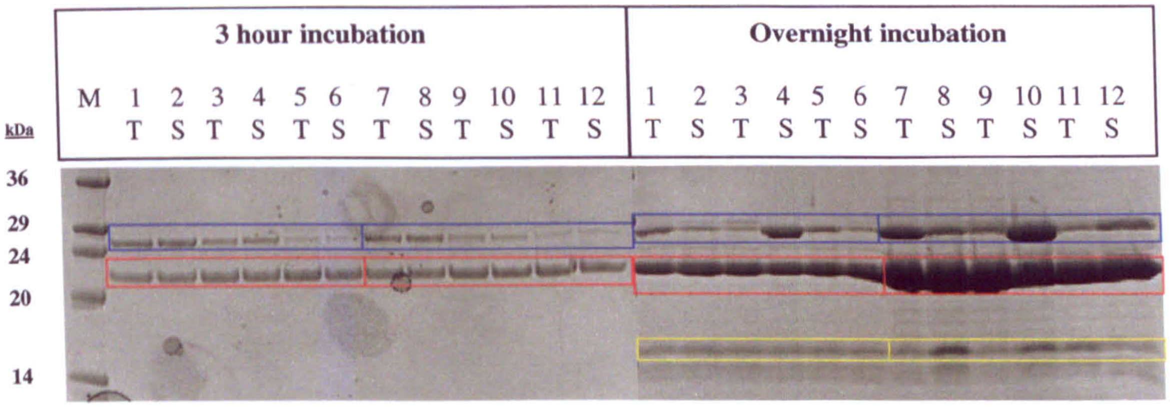


Figure 4.5: 10 % SDS-PAGE from cell-free *Tev* cleavage study of **Rv3628** (section 4.2.2f). Total extracts (T) and soluble fractions (S) from small-scale reactions, following affinity chromatography and incubation with *Tev* protease. Reactions incubated for the specified time at 4 °C with: **5 µg/ml *Tev* protease**, lanes 1 - 2; **10 µg/ml**, 3 - 4; **20 µg/ml**, 5 - 6. Reactions incubated for the specified time at 30 °C with: **5 µg/ml *Tev* protease**, lanes 7 - 8; **10 µg/ml**, 9 - 10; **20 µg/ml**, 11 – 12. Molecular weight markers in lane M. Target protein highlighted in blue (uncleaved) and red (cleaved). Cleaved His-tags highlighted in yellow (not visible in the 3 hour incubation wells, possibly due to a problem with sample loading).

4.2.3d Large scale protein expression and purification

Milligram quantities of soluble protein were obtained for all of the nine targets that were progressed into large-scale cell-free synthesis (**table 4.6**). All were synthesised in the presence of selenomethionine and without the inclusion of additives such as metal ions (with the exception of Rv3836, see section 4.2.2g). His-tags were removed by cleavage with *Tev* protease and affinity chromatography (**figure 4.7A**). Proteins were further purified to remove bulk contaminants (**table 4.6** and **figure 4.7B**). This was successful for all targets, with the exception of Rv2229c, which unexpectedly eluted from the gel filtration column in the void volume and was subsequently lost in the waste fraction. As the column separation range is 3 to 70 kDa and the Rv2229c monomer is 26.9 kDa, it seems likely that the protein formed an aggregate or an oligomeric structure. An example profile showing the purification of Rv3628 is shown in **figure 4.6**.

		Cell-free reaction parameters		Successfully purified[1]			Yield	
Target	**Function**	**Vol. (ml)**	**Expressed with additives**	**Affinity[2]**	**Anion**	**GF[3]**	**mg/ml reaction solution**	**Total (mg)**
Rv2229c	Conserved hypothetical protein	18	No	Yes	Yes	No*	n/a	n/a
Rv2234	Phosphotyrosine phosphatase	18	No	Yes	Yes	Yes	0.14	2.5
Rv2547	Transcriptional regulator protein	18	No	Yes	Yes	n/a	0.04	0.8
Rv2711	Iron-dependent repressor protein	18	No	Yes	Yes	Yes	0.05	0.9
Rv3042c	Phosphoserine phosphatase	18	No	Yes	Yes	Yes	0.28	5.0
Rv3628	Inorganic pyrophosphatase	18	No	Yes	Yes	Yes	1.49	26.9
Rv3836	Zinc metalloprotease	18	Yes (50 µm $ZnSO_4$)	Yes	Yes	n/a	0.33	6.0
Rv3867	Conserved hypothetical protein	18	No	Yes	Yes	n/a	0.18	3.2
Rv2981c	D-alanine-D-alanine ligase	18	No	Yes	Yes	n/a	0.11	2.0

Table 4.6: Results summary from the large-scale expression and purification of 9 *Mycobacterium tuberculosis* targets using the cell-free system. [1]Target judged to be ~ 95 % pure. [2]Affinity chromatography followed by removal of His-tags by *Tev* protease and affinity chromatography. [3]Proteins of inadequate purity were passed through a gel filtration column. *Rv2229c was lost during this step.

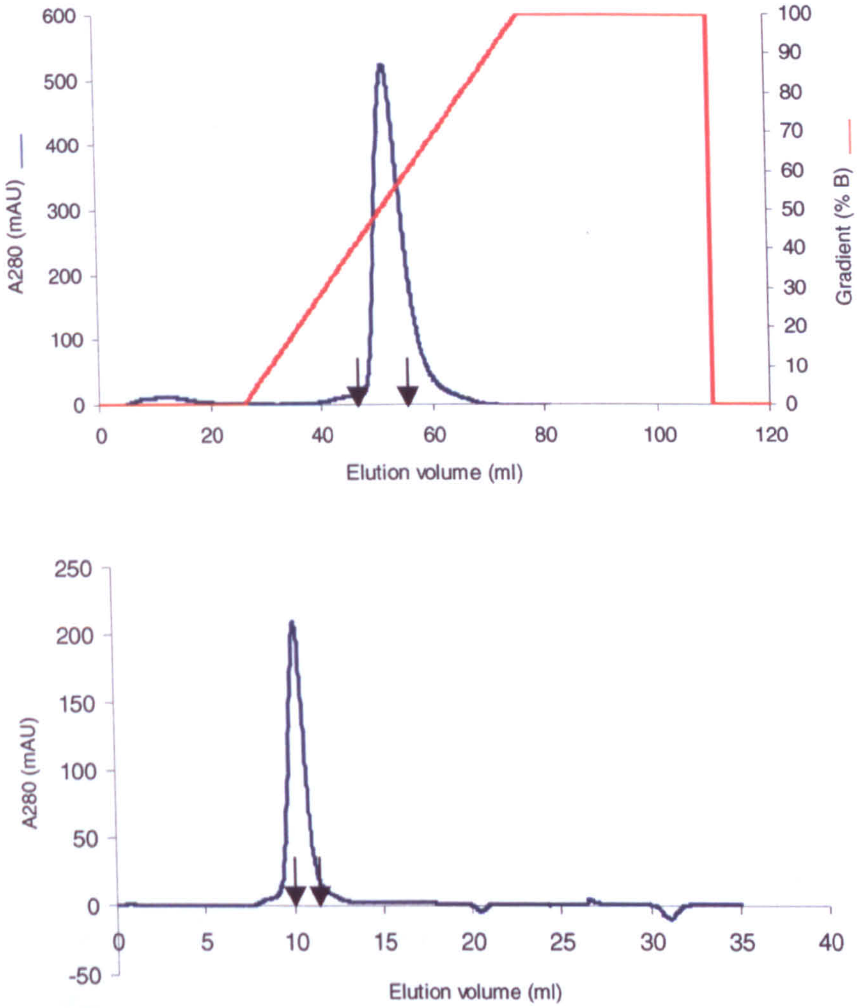
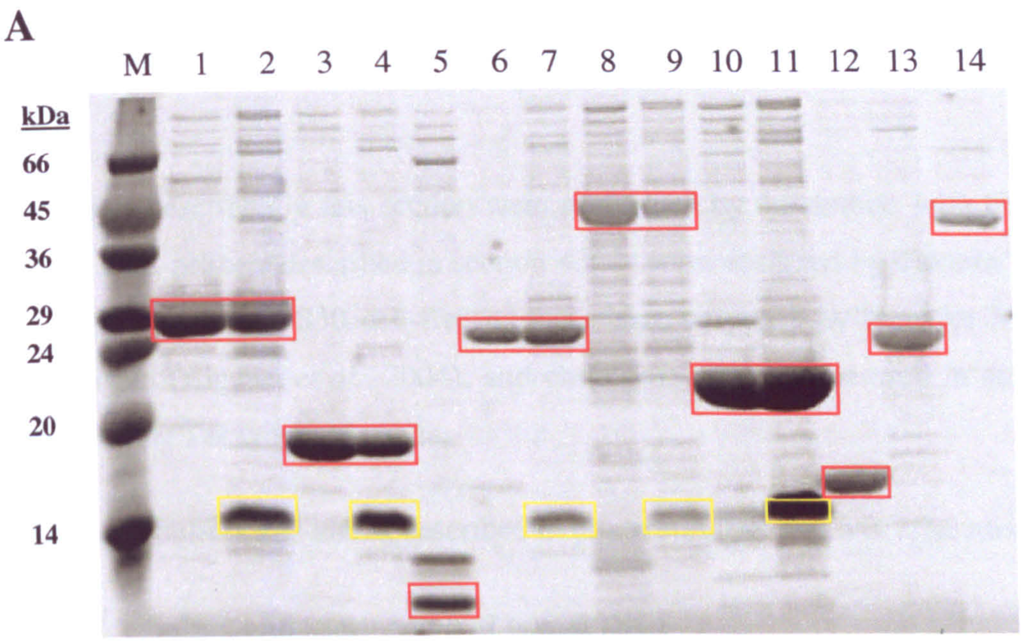


Figure 4.6: Example profiles showing the purification of Rv3628 by **(A)** anion exchange and **(B)** gel filtration chromatography, using an AKTA Explorer system (Amersham). Rv3628, synthesised using the cell-free expression system, was first applied to a Q sepharose anion exchange column (Amsersham) and eluted in a gradient of 0 – 100 % buffer CF-D. The peak fractions were pooled and applied to a S75 10/300 gel filtration column (Amersham), equilibrated in buffer CF-E. Pooled fractions are shown between arrows.



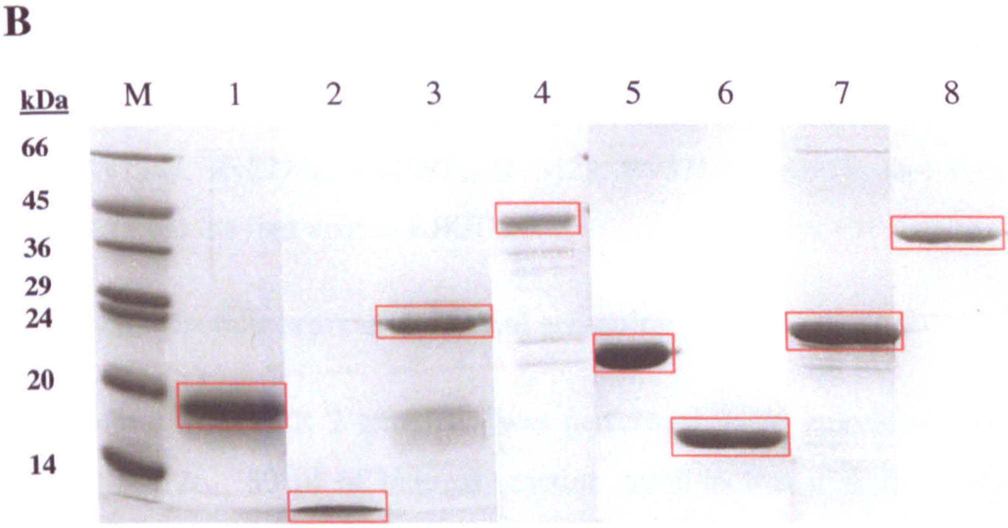


Figure 4.7: (A) 10 % SDS-PAGE from large-scale cell-free expression of nine targets (section 4.2.2g). Target proteins purified by affinity-chromatography and His-tags cleaved from proteins by incubation with *Tev* protease, followed by a further affinity chromatography step. **Rv2229c:** in lanes 1 (flowthrough) **(FT)** and 2 (elution) **(E)**. **Rv2234:** 3 **(FT)** and 4 **(E)**. **Rv2547:** 5 **(FT)**. **Rv2711:** 6 **(FT)** and 7 **(E)**. **Rv3042c:** 8 **(FT)** and 9 **(E)**. **Rv3628:** 10 **(FT)** and 11 **(E)**. **Rv3836** (synthesised with 50 μM $ZnSO_4$)**:** 12 **(FT)**. **Rv3867:** 13 **(FT)**. **Rv2981c:** 14 **(FT)**.

(B) 15 % SDS-PAGE of proteins after further purification (see **table 4.6**): **Rv2234,** 1; **Rv2547,** 2; **Rv2711,** 3; **Rv3042c,** 4; **Rv3628,** 5; **Rv3836,** 6; **Rv3867,** 7; **Rv2981c**, 8. Molecular weight markers in lanes M. Target proteins highlighted in red and cleaved His-tags in yellow.

4.2.4 Optimisation of expression conditions for 13 *Mycobacterium tuberculosis* targets: Methods

All procedures described in this section were performed by the author, with the following exceptions: PCR primers described in section 4.2.4a were designed by Takashi Yabuki and Yukiko Fujikura; *E. coli* S30 cell-free extracts were prepared by Natsuko Matsuda and Natsumi Suzuki (Kigawa *et al.*, 2004); and chaperone cell extracts used in section 4.2.4c were prepared by Takayoshi Matsuda.

The recipes for buffers and media described in this section are given in appendix 2.

4.2.4a PCR, cloning, and sequencing of target DNA

PCR and cloning steps were carried out as described in 4.2.2a to 4.2.2c. It was necessary to include DMSO in the PCR reactions for Rv0950c, Rv2388c, and Rv3534c, to obtain discrete bands visible by agarose gel electrophoresis. For targets Rv0185, Rv0247c, Rv2229c, Rv2547, Rv2776c, Rv2981c, Rv3628, Rv3717, Rv3915, and Rv3836, clones were retained from the first visit to RIKEN.

4.2.4b Cell-free protein expression: Initial screening

Initial screening from PCR 2 templates was performed using individual dialysis cups as described in 4.2.2e. 30 µl of internal reaction solution was dialysed against a 300 µl external solution at 30 °C for 6 hours. Initial screening was unnecessary for the repeat targets.

Expression and solubility were compared in the presence and absence of metal ions. 50 µM (final concentration) of zinc sulphate was added to the predicted zinc metalloproteins: Rv0670 and Rv2845c, and also to the unknown targets: Rv0950c, Rv3534c, and Rv3781. Iron sulphate was included for targets with iron-binding domains: Rv1589, Rv2388c, and Rv3545c, and also in the unknown proteins mentioned previously. Sequence homology database searches suggested that manganese and magnesium might bind to Rv3534c and Rv3781 respectively, so 50 µM of these metals (MnCl and ($MgCH_3CH_2)_2$) were included in the reaction solutions.

4.2.4c Cell-free protein expression: Optimisation

Optimisation of the cell-free reactions was performed as described in 4.2.2e, using plasmid templates, at 30 °C for 4 hours. This decrease in reaction incubation time was found to improve target protein solubility in some cases (Matsuda *et al.*, personal communication).

Detergents were included in the cell-free reaction solutions for targets which were not expressed in a soluble form during initial screening. With the help of Dr. Satoru Watanabe at RIKEN, the most favourable conditions were identified to be 0.5 to 1 % v/v of the non-ionic detergents Brij-35 (polyoxyethylene 23 lauryl ether, CMC 0.09 mM) or Digitonin (Ishihara *et al.*, 2004). These were included in both the internal and external reaction solutions at the above concentrations and also in the purification buffers at a final concentration of 0.01 % v/v (both Sigma), for those targets which were not expressed in a soluble form.

Finally, molecular chaperones were added during the cell-free synthesis of insoluble targets. Half of the *E. coli* S30 extract was substituted with an equal volume of *E. coli* extract containing a chaperone-plasmid construct. Five different chaperone combinations were chosen: 1. dnaJ-dnaK-grpE-groEL-groES; 2. groEL-groES; 3. dnaJ-dnaK-grpE; 4. groEL-groES-trigger factor; and 5. trigger factor (Matsuda *et al.*, personal communication). Due to the number of experimental conditions to be screened, the cell-free reaction was conducted in a 96-well micro-dialysis plate, as described in 4.2.2d.

Proteins which were solubilised by the inclusion of molecular chaperones, were purified by adding 30 µl of buffer CF-A to the internal reaction solution in a 96-well plate. 5 µl was removed for the total fraction and the plate centrifuged at 4, 500 rpm for 5 minutes at 4 °C. 5 µl of supernatant was removed for the soluble fraction. The supernatant, together with 40 µl of TALON^{TM} Superflow cobalt resin (BD Biosciences Clontech), were added to a 0.45 µM multi-screen-HV plate (Millipore) pre-equilibrated with buffer CF-A. Following a 20 minute incubation at room temperature, the plate was centrifuged at 1, 500 rpm for 1 minute at 4 °C, to obtain the flow-through fraction. Fractions were collected in a 96-well microtiter plate (ABgene). After three washes with 150 µl of buffer CF-A (wash fraction), His-tagged target proteins were eluted in 100 µl of buffer CF-B.

Target proteins which remained bound to molecular chaperones following affinity chromatography (Rv3717 and Rv3781), were incubated with 5 mM ATP solution (ATP-2NA, pH 7.0) for 30 minutes at 37 °C. A further affinity chromatography step was then performed, in an attempt to remove the chaperones.

4.2.4d Large scale protein expression & purification

Large scale expression and purifications were performed as described in 4.2.2g. Targets with low solubility were expressed during a shorter incubation period, to reduce the formation of aggregates (Matsuda *et al.*, personal communication). Additional large scale preparations were also performed for targets with particularly low yield. His-tags were cleaved from target proteins by incubation with *Tev* protease and purification by affinity chromatography (see section 4.2.2h).

Proteins were buffer exchanged into buffer CF-F, concentrated to 10 ml using Amicon 15 kDa cut-off centrifugal filters (Millipore) at 4, 000 x g, and loaded onto a 5 ml anion exchange column equilibrated in the same buffer (HiTrap^{TM} Q Sepharose^{TM} HP IEX, Amersham Biosciences). Target proteins were eluted in a gradient of 0 to 100 % buffer CF-G, desalted into buffer CF-H, and concentrated to 1.5 ml. 2 µl of each were loaded onto a 15 % polyacrylamide denaturing gel. Targets which were not pure enough for crystallisation (> 95 % purity) were concentrated to 1 to 1.5 ml and loaded onto a Superdex 75 10/300 (Amersham Biosciences) equilibrated in buffer CF-H. The remaining proteins were buffer exchanged into buffer CF-H.

Following purification steps, total yield was calculated by absorbance at 280 nm, using individual theoretical molar extinction coefficients (TBSGC). Bradford assays were not performed due to the limited quantities of protein available. Proteins were concentrated to 10 mg/ml, snap frozen in liquid nitrogen, and stored at –80 °C in 25 µl aliquots.

4.2.5 Optimisation of expression conditions for 13 *Mycobacterium tuberculosis* targets: Results

As described in section 4.2.3, only a selection of gel photographs are included in this section, additional SDS-PAGE photographs are included in appendix 1. A summary of results from small-scale expression trials is given in **table 4.7**.

4.2.5a Cloning of target DNA

PCR was used to successfully amplify single gene products of correct size for all targets. To achieve this it was necessary in include DMSO in the PCR reaction for targets Rv0950c, Rv2388c, and Rv3534c (see section 4.2.2a for a description of the use of DMSO during PCR). Positive clones of correct sequence were also obtained for all targets.

4.2.5b Cell-free protein expression: Initial screening

Some soluble expression was achieved without the addition of metals for Rv0670, Rv0950c, and Rv2845c. Rv1589 was soluble without metals and adding iron sulphate did not alter solubility. Solubility was marginally improved by the addition of zinc to Rv2845c, however no beneficial effect was observed when added to Rv0670. Rv0950c was also soluble in the presence of zinc ions but not iron, magnesium, or manganese. The remaining four targets were completely insoluble with or without addition of metal cofactors. Rv3534c was insoluble without additives, however was not expressed in the presence of manganese. This may be due to the metal ions affecting the cell-free reaction or due to an experimental error. Finally, the solubility of Rv3781 was not improved upon addition of magnesium, however an increase in total yield was observed, possibly due to the ions effect on RNA polymerase activity. See **figure 4.8**.

				Optimisation result		
Target	**Successful PCR (+ DMSO)[1]**	**Positive clone obtained[2]**	**Initial screening result[3]**	**Metals[4]**	**Detergents[5]**	**Molecular chaperones[6]**
Rv0185	Yes	Yes	Insoluble	Insoluble ($ZnSO_4$)	Soluble	Soluble
Rv0247c	Yes	Yes	Insoluble	Insoluble ($FeSO_4$)	Insoluble	Very low solubility
Rv2776c	Yes	Yes	Insoluble	Insoluble ($FeSO_4$)	Insoluble	Soluble
Rv3717	Yes	Yes	Insoluble	n/a	Very low solubility	Low solubility
Rv3915	Yes	Yes	Insoluble	n/a	Insoluble	Very low solubility
Rv0670	Yes	Yes	Soluble	Soluble ($ZnSO_4$)	n/a	n/a
Rv0950c	Yes (+)	Yes	Soluble	Soluble ($ZnSO_4$)	n/a	n/a
Rv1589	Yes	Yes	Soluble	Soluble ($FeSO_4$)	n/a	n/a
Rv2388c	Yes (+)	Yes	Insoluble	Insoluble ($FeSO_4$)	Soluble	Insoluble
Rv2845c	Yes	Yes	Soluble	Soluble ($ZnSO_4$)	n/a	n/a
Rv3534c	Yes (+)	Yes	Insoluble	No expression ($MnCl_2$)	Low solubility	Soluble
Rv3545c	Yes	Yes	Insoluble	Insoluble ($FeSO_4$)	Insoluble	Soluble
Rv3781	Yes	Yes	Insoluble	Insoluble ($((MgCH_3CH_2)_2)$)	Soluble	Soluble

Table 4.7: Results summary from the cloning and small-scale expression of 13 *Mycobacterium tuberculosis targets using the cell-free system.* [1]*Single PCR* products of correct size, synthesised in the presence or absence of DMSO (section 4.2.5a). [2]Clones free from nonsense, missense, and read-through mutations (section 4.2.5a). [3]Expression result from initial screening using standard conditions (section 4.2.5b). Expression result from optimisation of reaction conditions including: [4]Metal compounds (section 4.2.5b); [5]Detergents (section 4.2.5c); [6]Molecular chaperones (**table 4.8** and section 4.2.5c). Text in red describes results obtained from sections 4.2.3a to 4.2.3c.

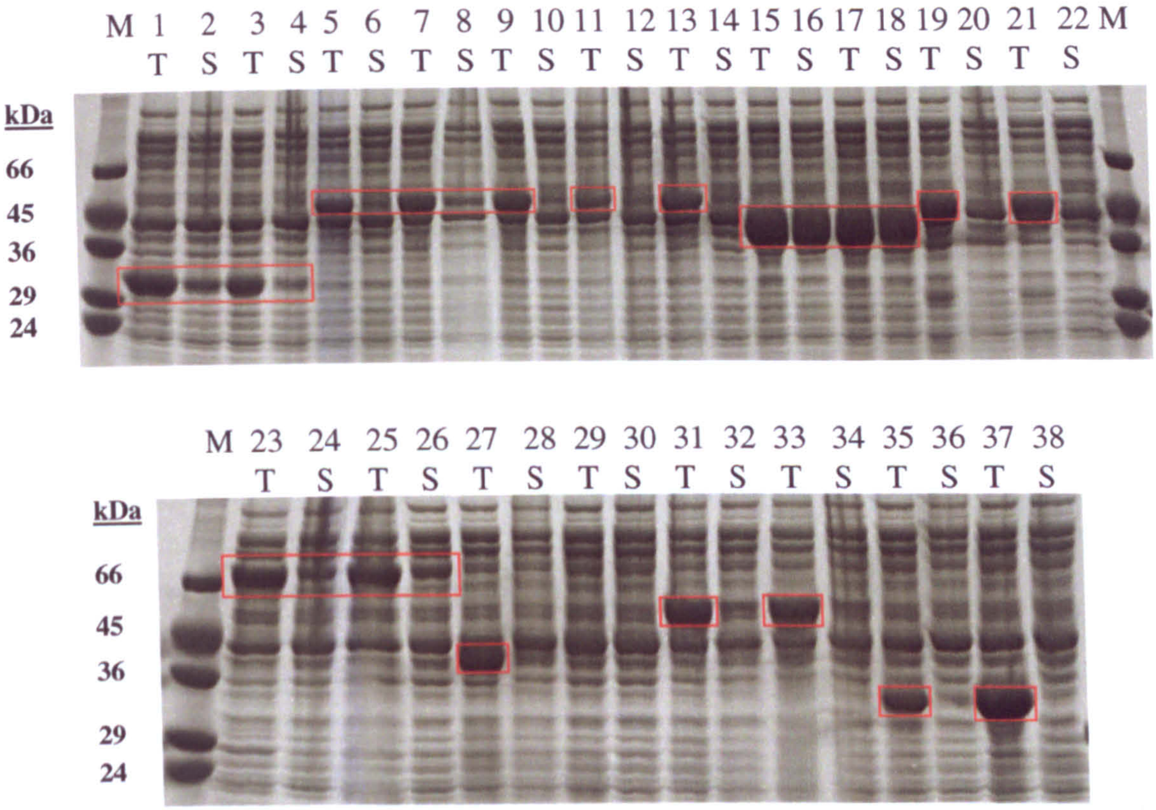

M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 M
T S T S T S T S T S T S T S T S T S T S T S
kDa
66
45
36
29
24
M 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38
T S T S T S T S T S T S T S T S
kDa
66
45
36
29
24

Figure 4.8: 10 % SDS-PAGE of initial cell-free screening, with and without the addition of metal compounds (50 μM) (section 4.2.4b). Total extracts (T) and soluble fractions (S) from reactions incubated at 30 °C for 6 hours. **Rv0670:** in lanes 1 – 2; with **$ZnSO_4$** 3 – 4. **Rv0950c:** 5 – 6; with **$ZnSO_4$** 7 – 8; **$FeSO_4$** 9 – 10; **$(MgCH_3CH_2)_2$** 11 – 12; **$MnCl_2$** 13 - 14. **Rv1589:** 15 – 16; with **$FeSO_4$** 17 – 18. **Rv2388c:** 19 – 20; with **$FeSO_4$** 21 - 22. **Rv2845c:** 23 – 24; with **$ZnSO_4$** 25 - 26. **Rv3534c:** 27 – 28; with **$MnCl_2$** 29 - 30. **Rv3545c:** 31 – 32; with **$FeSO_4$** 33 - 34. **Rv3781:** 35 – 36; with **$(MgCH_3CH_2)_2$** 37 - 38. Molecular weight markers in lanes M. Target proteins highlighted in red.

4.2.5c Cell-free protein expression: Optimisation

Addition of detergents

Addition of Brij-35 or Digitonin to reaction solutions increased solubility for targets Rv3781 (**figure 4.9**), Rv0185, Rv3534c, and Rv3717. Most noticeably, a marked increase in solubility was observed for Rv2388c upon addition of either detergent. However the inclusion of detergents had no beneficial effect during synthesis of Rv0247c, Rv2776c, Rv3545c, or Rv3915, and these targets remained insoluble. See **figures A3 and A4**, appendix 1.

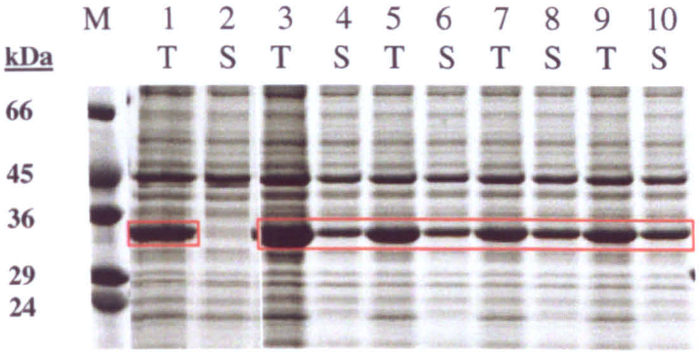


Figure 4.9: 10 % SDS-PAGE from optimisation of **Rv3781** cell-free expression conditions by the addition of detergents (section 4.2.4c). Total extracts (T) and soluble fractions (S) from small-scale reactions incubated at 30 °C for 4 hours. **No detergent** in lanes 1 – 2; **Brij-35**, 3 - 4 (0.5 %) and 5 - 6 (1 %); **Digitonin**, 7 - 8 (0.5 %) and 9 - 10 (1 %). Molecular weight marker in lane M. Rv3781 highlighted in red.

Molecular chaperones

Inclusion of molecular chaperones in the cell-free reaction solution proved highly successful in solubilising previously insoluble targets from initial screening (**table 4.8** and **figure 4.10**). Only one target, Rv2388c, was not expressed in a soluble form in the presence of chaperones. The most effective chaperone system, dnaJ-dnaK-grpE, improved solubility for eight targets. To remove molecular chaperones after synthesis, target proteins were purified by affinity chromatography, with an ATP incubation step when necessary, as described in section 4.2.4c. Rv0247c was not visible following affinity purification, suggesting aggregation or a misinterpretation of initial solubility. See also **figures A5 to A7**, appendix 1.

	Some soluble expression in the presence of:				
Target	**dnaK-dnaJ-grpE-groEL-groES**	**groEL-groES**	**dnaJ-dnaK-grpE**	**groEL-groES-trigger factor**	**Trigger factor**
Rv0185			Yes		
Rv0247c			Yes		
Rv2776c			Yes		
Rv3717			Yes (dnaJ)		
Rv3915		Yes (groEL)*	Yes		
Rv2388c					
Rv3534c			Yes		
Rv3545c		Yes (groEL)	Yes		
Rv3781			Yes (dnaJ)		

Table 4.8: Optimisation of the cell-free reactions by addition of molecular chaperones to insoluble targets from initial screening. Text in parentheses represents chaperones which remained bound to target protein, following affinity chromatography, and those marked with an asterisk represent chaperones which were successfully removed by incubation with ATP.

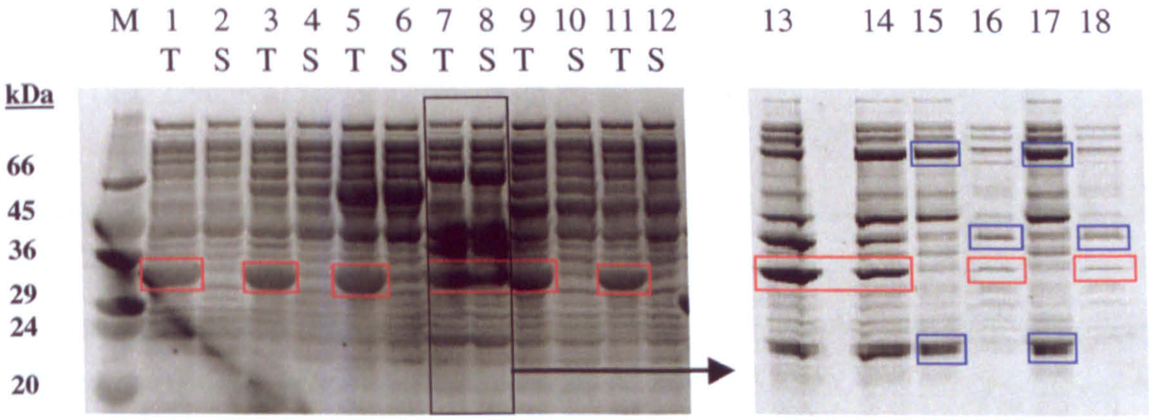


Figure 4.10: 10 % SDS-PAGE from optimisation of **Rv3781** cell-free expression conditions by the addition of molecular chaperones (section 4.2.4c). Total extracts (T) and soluble fractions (S) from small-scale reactions incubated at 30° C for 4 hours. **No chaperones** in lanes 1 – 2; with **dnaJ-dnaK-grpE-groEL-groES**, 3 – 4; **groEL-groES**, 5 – 6; **dnaJ-dnaK-grpE**, 7 – 8; **groEL-groES-trigger factor**, 9 – 10; **trigger factor**, 11 - 12. Fractions from affinity chromatography of **Rv3781** synthesised in the presence of dnaJ-dnaK-grpE: **Total**, 13; **Soluble**, 14; **Flow-through before ATP incubation**, 15; **Elution before ATP incubation**, 16; **Flow-through after ATP incubation**, 17; **Elution after ATP incubation**, 18. Molecular weight markers in lane M. Rv3781 highlighted in red and chaperones in blue.

Selenomethionine incorporation

No decrease in solubility occurred when synthesising Rv0670, Rv1589, or Rv3781 with selenomethionine, however no target protein was visible in the elution fraction following affinity-chromatography, by SDS-PAGE, for Rv0950c. It was not necessary to determine the effect of selenomethionine on solubility for the 8 repeat targets, as this was shown to have no effect during previous experiments (see section 4.2.3c).

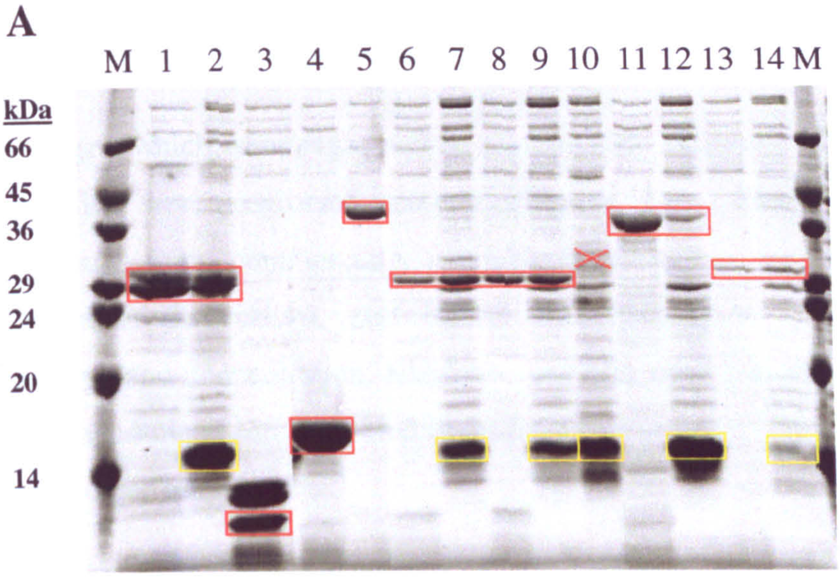
4.2.5d Large scale protein expression and purification

Four soluble targets were selected for large scale synthesis (Rv0670, Rv0950c, Rv1589, and Rv3781), to produce sufficient quantities of protein for crystallisation trials (**table 4.9** and **figure 4.11A**). Those targets which required molecular chaperones for solubility were discarded due to the time required to prepare sufficient quantities of chaperone extract and the potential problems associated with removing chaperones for downstream applications (**table 4.8**). Targets for which the addition of Digitonin was essential for soluble expression were also discarded due to the high cost of the detergent. Finally, those which resulted in very low target yield were not included in large-scale synthesis. Large scale preparations for the four targets which were successfully expressed previously, section 4.2.3d (Rv2229c, Rv2547, Rv2981c, and Rv3836), were also performed (**figure 4.11A**).

Following purification by affinity chromarography, His-tags were successfully removed from targets by incubation with *Tev* protease, followed by an additional affinity chromatography step (**figure 4.11A**). Rv0950c was not visible by SDS-PAGE following affinity chromatography, possibly due to aggregation, and so was not progressed further. Rv3781 was also discarded due to its extremely low yield (< 0.5 mg, as estimated by SDS-PAGE), which would be unsuitable for crystallisation trials as further purification steps were required. All targets with the exception of Rv0670 and Rv1589 were considered pure enough for crystallisation (> 95 % pure by SDS-PAGE) after an anion exchange chromatography step (**figure 4.11B**). The purity of Rv1589 was improved by gel filtration chromatography, however Rv0670 was co-purified with a number of contaminants following this additional step. Further purification of this target was not performed due to the low yield of protein available. Rv2547 was co-purified with an unknown protein of ~ 10 kDa.

		Cell-free reaction parameters				Yield	
Target	**Function**	**Vol. (ml)**	**30 °C incubation time (hr)**	**Expressed with additives**	**Successfully purified[1]**	**mg/ml reaction solution**	**Total (mg)**
Rv0670	Endonuclease IV	144	4	Yes ($ZnSO_4$)	Partially (larger MW contaminants remained)	0.03	4.8
Rv0950c	Hypothetical metalloprotease	72	4	No	No (possible degradation)	n/a	n/a
Rv1589	Biotin synthetase	72	4	No	Yes	0.11	7.56
Rv3781	ATP-binding protein ABC transporter	72	4	Yes (0.5 % v/v Brij-35)	No (yield too low)	n/a	n/a
Rv2229c	Conserved hypothetical protein	18	8	No	Yes	0.18	3.2
Rv2547	Transcriptional regulator	18	8	No	Yes	0.07	1.2
Rv2981c	D-alanine-D-alanine ligase	18	8	No	Yes	0.04	0.8
Rv3836	Zinc metalloprotease	18	8	Yes ($ZnSO_4$)	Yes	0.35	6.3

Table 4.9: Results summary for the large-scale expression and purification of 8 *Mycobacterium tuberculosis* targets using the cell-free system. [1]Target proteins ~ 95 % pure by: 1. Affinity chromatography; 2. Incubation with *Tev* protease followed by an additional affinity chromatography step, to remove His-tags; and 3. Anion exchange chromatography. A final gel filtration step was also performed for Rv0670 and Rv1589.



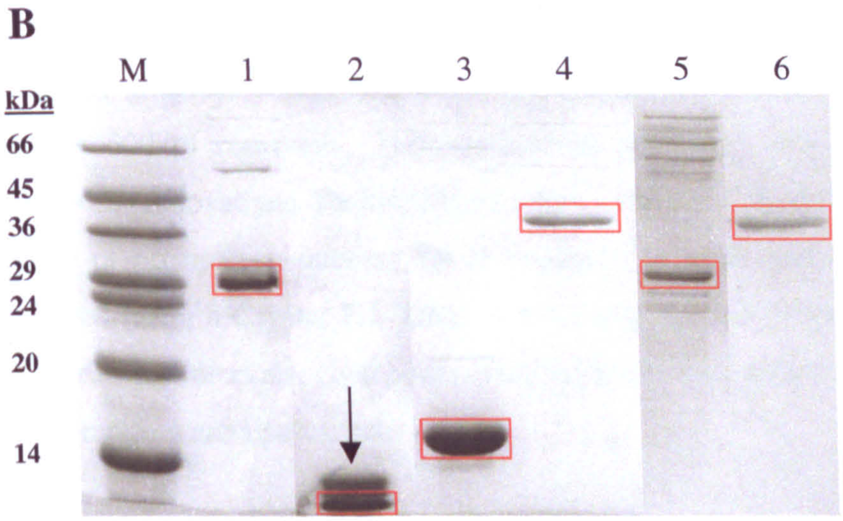


Figure 4.11: (A) 10 % SDS-PAGE from large-scale cell-free expression of eight targets (section 4.2.4d). Target proteins purified by affinity chromatography and His-tags cleaved from proteins by incubation with *Tev* protease, followed by a further affinity chromatography step. **Rv2229c:** in lanes 1 (flow-through) (**FT**) and 2 (elution) (**E**). **Rv2547:** 3 (**FT**). **Rv3836** (synthesised with 50 μM $ZnSO_4$)**:** 4 (**FT**). **Rv2981c:** 5 (**FT**). **Rv0670** (synthesised with 50 μM $ZnSO_4$)**:** 6 & 8 (**FT**) and 7 & 9 (**E**). **Rv0950c:** 10 (not visible following affinity-chromatography) (**FT**). **Rv1589:** 11 (**FT**) and 12 (**E**). **Rv3781** (synthesised with 0.5 % Brij-35)**:** 13 (**FT**) and 14 (**E**).

(B) 15 % SDS-PAGE of proteins after further purification by anion exchange chromatography: **Rv2229c,** 1; **Rv2547,** 2 (the co-purified band is shown with an arrow); **Rv3836,** 3; **Rv2981c,** 4. Fractions after further purification by anion exchange and gel filtration chromatography: **Rv0670,** 5; **Rv1589,** 6. Molecular weight markers in lanes M. Target proteins highlighted in red and cleaved His-tags in yellow.

4.2.6 Crystallisation trials of target proteins

All ten of the targets which were expressed on a large-scale using the cell-free system and successfully purified, were progressed into crystallisation trials. Both manual and robotic trials was performed using commercially available broad matrix screens from Hampton Research, Molecular Dimensions, and Nextal Biotechnologies. Variables such as salt/precipitant type and concentration, additives, and pH, were screened using these kits. Where possible, hand-made screens based upon conditions used to crystallise homologous proteins, were also performed.

Manual screening was performed using a standard 24-well pre-greased plate suitable for hanging-drop vapour-diffusion crystallisation (VDX plate, Hampton Research). 1 to 2 µl of protein was mixed with an equal volume of precipitant on a siliconised cover slip, suspended over a 500 µl reservoir. High-throughput screening was achieved using a Screenmaker 96 + 8 (Innovadyne Technologies) robot. 100 nl of protein was mixed with an equal volume of precipitant, over an 80 µl reservoir, in a 96-well sitting-drop plate. Plates were viewed using a Crystal Pro robot with Crystal L.I.M.S. software (both Tritek Corporation), at regular intervals. For both screening methods, protein concentration and incubation temperature were also varied.

Despite screening each target against hundreds of unique conditions, only one protein gave crystals suitable for X-ray diffraction. The 2.7 Å crystal structure of Rv3628, an inorganic pyrophosphatase, is described in chapter 5. This low success rate may partly be due to sample heterogeneity, resulting from extended incubations at 4 °C during shipment of the samples from Japan to England.

4.3 *In vivo* protein expression

4.3.1 Introduction

To enable comparisons to be drawn between the cell-free system and more traditional expression systems, a number of targets described in section 4.1 were selected for expression trials using an *E. coli in vivo* system. Four targets were selected based upon their outcome from the cell-free expression trials. Rv3628 was selected for its high yield and solubility and Rv3836 for its reduced, but soluble yield. Rv0950c was chosen because it appeared soluble throughout expression trials, however was not detectable following affinity purification, possibly due to aggregation. Finally, Rv3545c was chosen as it remained completely insoluble, except in the presence of molecular chaperones. This target, a cytochrome P450 125, was also selected due to our group's ongoing interest in cytochromes.

Oligonucleotide primers were designed from the genomic sequence of each target and were amplified by PCR. PCR products were inserted into a vector and were used to transform a number of different *E. coli* host strains, from which small-scale expression cultures were set up. Expression parameters were altered to obtain optimal yields of soluble protein before progressing, when appropriate, into large-scale production. Rv3545c was purified by successive chromatographic steps and used for downstream applications (see chapter 6).

The cloning and expression of Rv3545c was based upon current literature of homologous cytochrome P450s from *Mycobacterium tuberculosis* and was performed separately from the remaining three targets.

4.3.2 Cloning and expression of Rv3545c: Methods

The recipes for buffers and media described in this section are given in appendix 2.

4.3.2a Rv3545c PCR

The Rv3545c gene was amplified directly from the genome using a standard hot-start PCR protocol, from genomic *M. tb* DNA (H37Rv) obtained from Colorado State University. Oligonucleotide primers were designed based on those used to amplify a Rv3545c homologue, *M. tb* CYP51 (Bellamine *et al.*, 1999), to incorporate a 5' upstream *Nde*I restriction site, a C-terminal 4-His tag, and a downstream *Hin*dIII site (**figure 4.12**). To improve restriction enzyme recognition, a 3-nucleotide linker sequence was incorporated, flanking each of the restriction sites. 1.2 μg of genomic DNA was included in a typical 80 μl PCR reaction solution of: 50 nM of each primer (Operon); 0.2 mM of each dNTP (dATP, dCTP, dGTP, and dUTP) (Novagen); 2 U Vent$_R$ HiFi DNA polymerase (New England Biolabs); and 1 x HiFi buffer (New England Biolabs). An initial denaturation of double-stranded DNA at 94 °C for 5 minutes was followed by 30 cycles of 94 °C for 30 seconds, followed by annealing of the primers at 62 °C for 30 seconds, and an extension period of 72 °C for 45 seconds. A final extension at 72 °C for 10 minutes completed the cycle. PCR reactions were performed using a GeneAmp® PCR System 2700 (Applied Biosystems).

The PCR product was visualised by electrophoresis on a gel containing 1 % agarose and 0.5 μg/ml of ethidium bromide (Sigma) in 1 x TBE buffer. The gel was run at an electrical potential of 150 V and viewed briefly under ultraviolet fluorescence. The PCR product of correct size was excised from the gel using a clean blade and purified using a QIAquick Gel Extraction kit (Qiagen), according to manufacturer's instructions. Purified DNA was eluted in 50 μl of autoclaved MilliQ water, before storing at – 20 °C.

FW 5' CGCCATATGTCGTGGAATCACCAGTCA

RV 5' CGCAAGCTTCAGTGATGGTGATGGTGAGCAACCGGGCATCTACCGG

Figure 4.12: Primer sequences used to amplify the Rv3545c gene for *in vivo* expression. Bold text identifies linker sequences (blue), restriction sites (red), start and stop codons (underlined), and the C-terminal His-tag (grey). Normal text identifies Rv3545c-unique sequences.

4.3.2b Cloning of Rv3545c

The Rv3545c PCR product was directly cloned into the expression vector pET17b (Novagen). Both the PCR product and pET17b were digested with *Nde*I and *Hin*dIII to produce compatible sticky ends for cloning. In separate 1.5 ml Eppendorf tubes, 20 μl of gel-purified PCR product and 3 μg of pET17b plasmid DNA were mixed with 20 U of each restriction enzyme and 1 x reaction buffer #2 (all New England Biolabs) in 30 μl reactions. Reactions were incubated at 37 °C for 3 hours. The digested pET17b plasmid was dephosphorylated to prevent recircularisation during the ligation reaction by incubating with 0.2 U (~ 0.05 U per pmol of DNA ends) of calf intestinal alkaline phosphatase (Roche) for a further 30 minutes. Digests were viewed by agarose gel electrophoresis and correctly sized bands were purified using the QIAquick Gel Extraction kit (Qiagen). The Rv3545c fragment and the linearised pET17b were eluted in 20 μl and 30 μl of MilliQ water, respectively.

12 µl of Rv3545c was ligated into 100 ng of pET17b in a 20 µl reaction containing 1 U T4 DNA ligase and 1 x ligase buffer (both Roche) by incubating overnight at 15 °C. Novablue single cells (Novagen) were transformed with ligation reaction, using the standard heat-shock protocol described by Novagen (Novagen, 2006). Cells were thawed on ice, mixed with 1 µl of ligation reaction, and incubated on ice for 5 minutes. Cells were heat-shocked in a 42 °C waterbath for 30 seconds and returned to ice for a further 2 minutes. 250 µl of SOC medium was added and 25 to 50 µl of cells were plated onto LB-Amp agar plates containing 100 µg/ml ampicillin. Plates were incubated overnight at 37 °C.

Sixteen single clones were used to independently inoculate 5 ml of LB-Amp media. Cultures were incubated overnight at 37 °C and plasmid DNA was purified using a QIAprep Spin Miniprep kit (Qiagen). Glycerol stocks were first made by removing 900 µl of each culture and mixing with 100 µl of 80 % glycerol, before freezing at -80 °C. The Rv3545c-pET17b plasmids were purified from 3 ml of culture, following manufacturers instructions, and eluted in 100 µl of MilliQ water, before storing at – 20 °C.

Initial screening for positive clones was performed by digesting the purified plasmids (7.5 µl) with *Nde*I and *Hin*dIII, as described previously. This enabled the identification of clones containing both pET17b and the Rv3545c insert but did not take into account mutations which may have occurred during PCR. Five of these clones were sent to Oxford University for DNA sequencing, using an Applied Biosystems 9700 Thermal Cycler (http://polaris.bioch.ox.ac.uk/dnaseq). Sequence files were aligned with the correct Rv3545c sequence, obtained from the TBSGC, using ClustalW (Thompson *et al.*, 1994), and positive clones were identified.

One positive clone was used to transform HMS174 (DE3) cells (Novagen), a host used to successfully express the homologous *M. tb* cytochrome P450, CYP51 Rv0764c (Bellamine *et al.*, 1999). The same clone was also used to transform Rosetta 2 (DE3) cells (Novagen), a host used to express proteins which utilise non-*E. coli* codons (Novagen, 2006). 20 µl of cells were transformed with 1 µl of purified plasmid, as described before. In each case, a single colony was grown overnight at 37 °C in 5 ml of LB-Amp media, for HMS174 (DE3) cells, and LB-Amp-Cam (including 34 µg/ml chloramphenicol) media, for Rosetta 2 (DE3) cells. 900 μl of overnight culture was removed for a glycerol stock, as described previously.

4.3.2c Expression trials of Rv3545c: Method 1

The expression and extraction protocols detailed here are modified from those described by Bellamine *et al.* (1999). A starter culture of Rv3545c-HMS174 (DE3) was obtained by inoculating 5 ml of LB-Amp with 3 μl of glycerol stock and incubating overnight at 37 °C in a 50 ml falcon tube.

Overnight culture was diluted ten fold in fresh TB-Amp and grown to an OD 600 of 0.6 – 1.0. The haem precursor, δ-aminolevulinic acid (Sigma), was added to a final concentration of 2 mM and cultures were induced with 1 mM (final concentration) of IPTG, before incubating for the specified time and temperature, whilst shaking at 185 rpm (**table 4.10**). The OD 600 was recorded, before pelleting cells by centrifugation at 8, 000 rpm in a Beckman JA20 rotor for 20 minutes, and pellets were stored at – 20 °C overnight.

To obtain the soluble fraction, thawed cells were resuspended in 0.125 x volume of TES buffer and incubated on a stirring platform with 0.5 mg/ml lysozyme (Sigma) for 15 minutes at 4 °C. One volume of 0.1 mM EDTA (Sigma) solution was added and then incubated for a further 30 minutes, before pelleting the spheroplasts at 3, 000 x g for 20 minutes. The supernatant was incubated with 1 μg/ml of DNaseI (Sigma) at 4 °C and ultracentrifuged at 225, 000 x g for 30 minutes in a Beckman 70Ti rotor. The resulting supernatant (supernatant A) was stored at 4 °C. Spheroplasts were resuspended in two-fold diluted TES buffer and sonicated at a 50 % output for 5 cycles of 30 second bursts, each followed by a 1 minute recovery period, on ice using a Branson sonicator. Lysates were ultracentrifuged, as before. No brown/red colour, indicative of the presence of P450, was visible in the periplasmic fraction and so supernatant A was combined with supernatant B to form the soluble fraction. All centrifugation steps were performed at 4 °C. Fractions were normalised for SDS-PAGE using **equation 4.1** (Novagen, 2006).

Total fractions were prepared by pelleting 1 ml of cell culture in a bench-top Eppendorf centrifuge for 1 minute at 14, 000 rpm and resuspending in 100 μl of 1 x PBS buffer. Cells were sonicated for several seconds at 30 % output and then denatured with 150 µl of 1 x sample loading buffer at 90 °C for 5 minutes.

Normalised soluble fractions were denatured at 90 °C for 5 minutes, in 1 x sample loading buffer. Samples were applied to a 15 % polyacrylamide gel in 1 x tris-glycine running buffer and run at 35 mA for 30 minutes, before viewing under white light. Expression conditions are described in **table 4.10**.

Equation 4.1:

$$Z = OD\ 600 \times DF$$

$$V = 80\ \mu l / Z$$

Where:

Z = undiluted OD 600 reading

DF = dilution factor of the sample

V = normalised volume in microlitres of sample

	Incubation Parameters				
Expression variable	**Time (hours)**	**Temp. (°C)**	**Media**[1]	**Additives**	**Reason**
Initial Screen	6 24	30	TB	None	Initial conditions
Temp.	6	18 25 30	TB	None	Lower incubation temperatures can improve protein solubility (Novagen, 2006)
Glucose	24	25	TB	0.5 % glucose added to 5 ml overnight starter culture	Decrease any basal expression, to prevent plasmid instability (Grossman *et al.*, 1998)

Table 4.10: Small-scale (50 ml) optimisation of *in vivo* expression conditions for Rv3545c in pET17b and HMS174 (DE3). [1] Terrific broth containing 100 µg/ml ampicillin.

4.3.2d Expression trials of Rv3545c: Method 2

The following method is a modification of that used to successfully express the human P450, CYP2C9 (Williams *et al.*, personal communication). Rv3545c (Rosetta 2) starter cultures (LB-Amp-Cam) were prepared and diluted in fresh TB-Amp-Cam, following the method described in section 4.3.2c. 1L cultures were grown to an OD 600 of 0.35 to 0.45, before adding 80 mg of δ-aminolevulinic acid (Sigma). Following a further incubation at 37 °C (for approximately 30 minutes), cultures were induced at an OD 600 of 0.7 - 0.8 with 1mM of IPTG (final concentration) and incubated at 25 °C for 24 and 72 hours, with shaking at 185 rpm. The OD 600 was recorded and cells were pelleted by centrifugation at 6, 000 rpm in a Beckman JLA8.1 rotor for 15 minutes and snap-frozen in liquid nitrogen, before storing at – 80 °C overnight.

Thawed cells were resuspended in 125-lysis buffer (10 ml per litre of culture) and disrupted by passage three times through a French pressure cell at 10, 000 pounds per square inch. The soluble fraction was obtained by centrifugation at 14, 000 rpm in a Beckman JA20 rotor for 30 minutes. All centrifugation steps were performed at 4 °C. Fractions were normalised for SDS-PAGE as described in section 4.3.2c.

4.3.2e Large scale expression and purification of soluble Rv3545c

Rv3545c was expressed and the soluble fraction obtained as outlined in section 4.3.2d. 6 L of Rosetta 2 (DE3) culture was pelleted and resuspended in 50 ml of 125-Lysis buffer, and the soluble fraction was incubated overnight with 15 ml of Nickel Sepharose High Performance resin (Amersham) at 4 °C on a rolling platform. The slurry was applied to a 100 ml empty column (Amersham), attached to a downstream peristaltic pump, and the resin was washed with 10 column volumes (150 ml) each of buffers 125-NiA and 125-NiB. Bound protein was eluted in 2 column volumes of buffer 125-NiC and dark red/brown fractions were pooled (10 ml).

Protein was concentrated to 4 ml in a Vivaspin-20 30 kDa MWCO filtration unit by centrifugation at 3, 500 x g. The concentrated sample was loaded onto a Superdex Hi-Prep S200 26/60 gel filtration column (Amersham) in two applications. The column, attached to an automated ÄKTA Explorer platform (Amersham), was equilibrated in 2 column volumes (640 ml) of 125-GF buffer. The sample (2ml) was applied to the column and eluted in 1.2 column volumes of the same buffer, at a flow rate of 3 ml/min with a maximum pressure limit of 0.5 MPa. Absorption measurements were recorded at 280 nm and 392 nm (the Soret peak of high-spin P450) and peaks were collected in 1 ml fractions at an absorbance (280 nm) above 200 mAu. Nine fractions from the centre of the peak were collected (18 ml total) and concentrated to 40 mg/ml, as described previously (concentration determined by Bradford assay).

4.3.3 Cloning and expression of Rv3545c: Results

Amplification of the Rv3545c by PCR and subsequent cloning steps were successful. Rv3545c production was observed when following the initial conditions described in method 1 (section 4.3.2c), however solubility was negligible (**figure 4.13A**). Reducing the incubation temperature to 25 °C did little to increase solubility, however a marginal improvement was observed at 18 °C (**figure 4.13B**). The addition of glucose during the overnight incubation, with the aim of stabilising the plasmid by decreasing basal expression, had no beneficial effect (**figure 4.13A**).

Significant yields of soluble protein were only achieved when expression incubations were extended to 72 hours, as described in method 2 (section 4.3.2d). Dark red/brown pellets (**figure 4.15A**), obtained from Rosetta 2 (DE3) cultures incubated at 25 °C, were suggestive of soluble P450 expression and this was confirmed by spectrometric analysis (section 6.3). Large-scale expression of Rv3545c under these conditions (section 4.3.2e) produced significant quantities of protein with a Soret peak characteristic of cytochrome P450s (in a low-spin haem-iron system) present at 426 nm, after purification by affinity chromatography (section 6.3.2a). Further purification by gel filtration yielded 13 mg of pure protein per litre of culture (80 mg total) using this procedure, with a Soret peak at 392 nm (high-spin system). See **figures 4.14B** and **4.15** for the large-scale purification of Rv3545c.

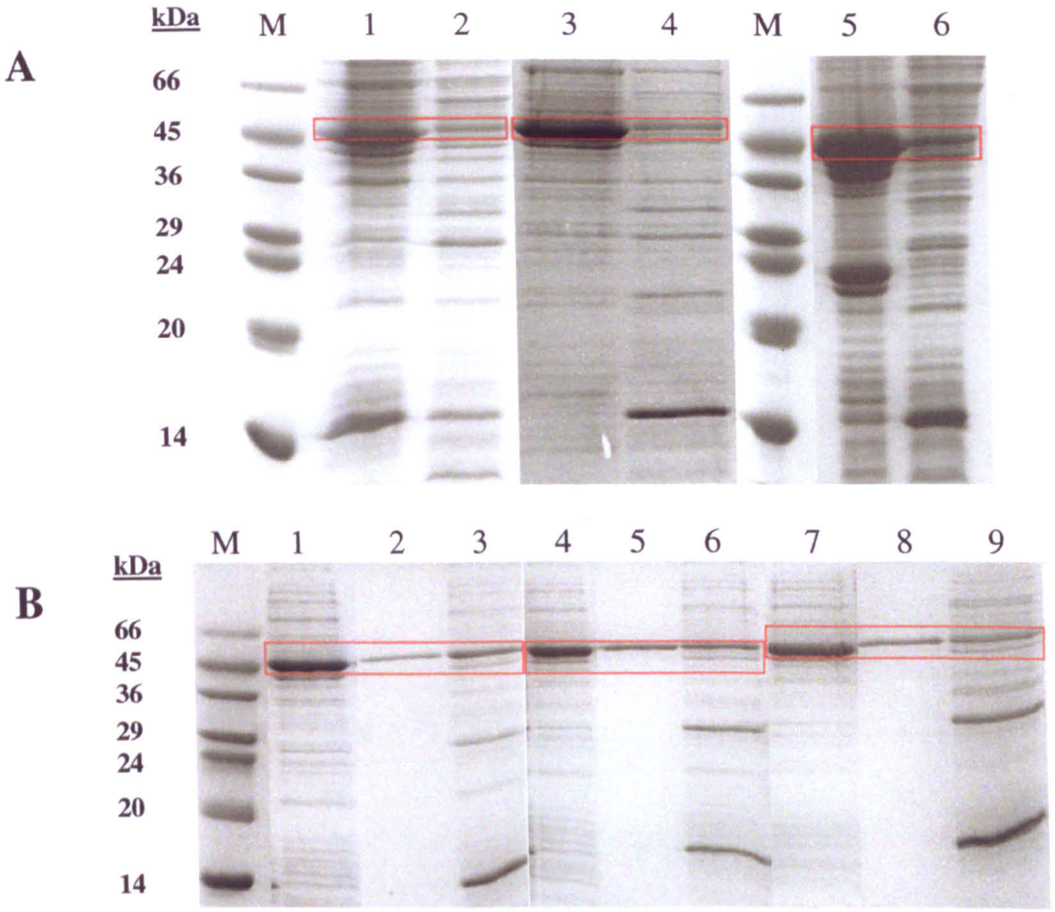


Figure 4.13: **(A)** 15 % SDS-PAGE from small-scale *in vivo* expression trials of Rv3545c in pET17b and HMS174 (DE3) (section 4.3.2c). Cultures incubated at 30 °C for: **6 hours**, in lanes 1 (insoluble) and 2 (soluble); **24 hours**, 3 (insoluble) and 4 (soluble). Cultures grown at 25 °C for 24 hours from starter cultures containing **0.5 % glucose**: 5 (insoluble); 6 (soluble).

(B) Cultures incubated for 6 hours at various temperatures. **18 °C:** total, 1; insoluble, 2; soluble. 3. **25 °C:** total, 4; insoluble, 5; soluble, 6. **30 °C:** total, 7; insoluble, 8; soluble, 9. Molecular weight markers in lanes M. Rv3545c highlighted in red.

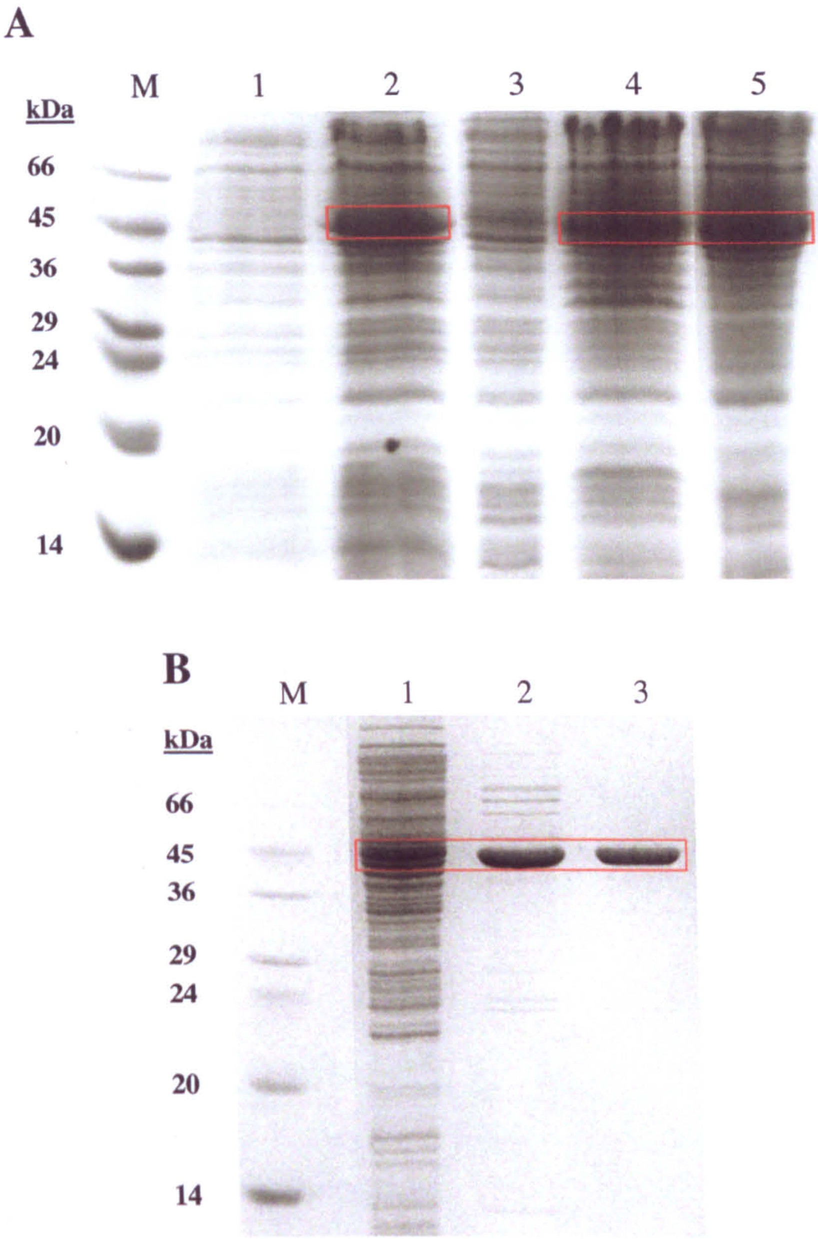


Figure 4.14: **(A)** 15 % SDS-PAGE from small-scale *in vivo* expression trials of Rv3545c in pET17b and Rosetta 2 (DE3) (section 4.3.2d). Cultures incubated at 25° C for 24 and 72 hours. **Uninduced** in lane 1; **total** and **soluble**, 2 - 3 (**24 hours**) and **total** and **soluble**, 4 - 5 (**72 hours**).

(B) Fractions from purification of Rv3545c, following large-scale production and purification (section 4.3.2e). Cultures incubated at 25 °C for 72 hours. **Soluble lysate** in lane 1; **post-affinity chromatography**, 2; **post-gel filtration**, 3. Molecular weight markers in lanes M. Rv3545c highlighted in red.

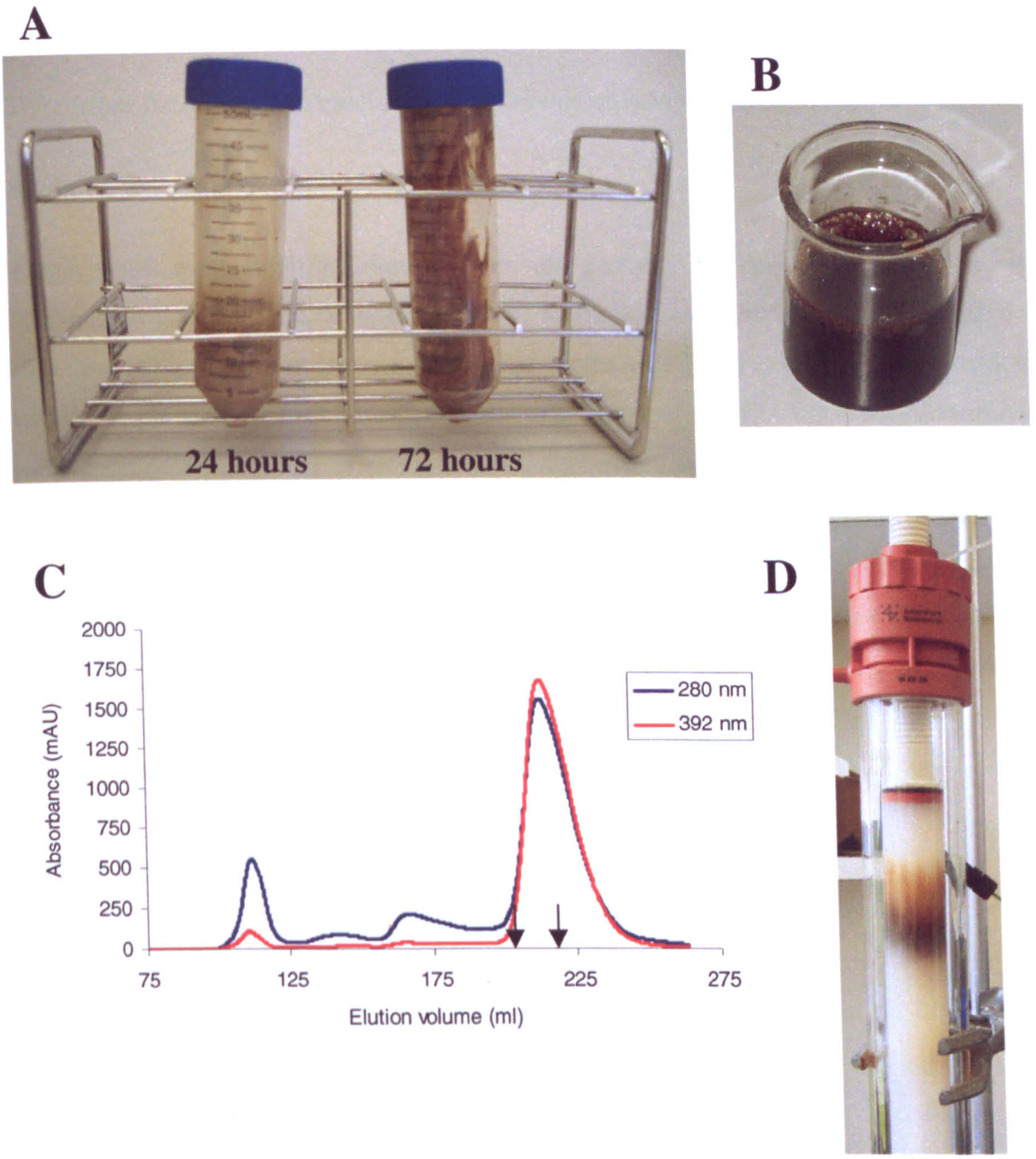


Figure 4.15: Large-scale *in vivo* expression and purification of Rv3545c.

(A) Cell pellets from Rv3545c expressed in Rosetta 2 (DE3) for 24 and 72 hours at 25 °C (section 4.3.2d). Both yield and intensity of the brown/red colour (indicative of haem proteins) are improved with prolonged expression incubations.

(B) Rv3545c after cell lysis and purification by affinity chromatography. Rosetta 2 (DE3) cultures (6 L) grown at 25 °C for 72 hours (section 4.3.2e).

(C) Profile from the purification of Rv3545c by gel filtration using an S200 26/60 column (Amersham Biosciences), from Rosetta 2 (DE3) culture (3 L) grown at 25 °C for 72 hours. Absorbance at 392 nm is predominantly due to the presence of cytochrome P450 in a high-spin system. Pooled fractions are shown between arrows (section 4.3.2e).

(D) Purification of Rv3545c by gel filtration, as before.

4.3.4 Cloning and expression of Rv0950c, Rv3628, and Rv3836: Methods

The recipes for buffers and media described in this section are given in appendix 2.

4.3.4a PCR

Target genes were amplified directly from the genome. Oligonucleotide primers were designed to incorporate a 5' upstream *Nde*I restriction site and a downstream *Xho*I site (**figure 4.16**) for insertion into the expression vector, pET28a (Novagen). This positioning within the pET28a construct resulted in the expression of a 6-His tag at the N-terminal of each target protein. 1.2 µg of genomic DNA was included in a typical 80 µl PCR reaction solution of: 50 nM of each primer (Operon); 0.2 mM of each dNTP (dATP, dCTP, dGTP, and dUTP) (Novagen); 2 U *Taq* DNA polymrease in storage buffer B; and 1 x *Taq* buffer including 15 mM $MgCl_2$ (both Promega). PCR reaction conditions were as described in section 4.3.2a, with the exception of the annealing temperature which was decreased to 58 °C. This was necessary due to the low Tm (64 °C) of the Rv3628_Rv primer. PCR products were purified from an agarose gel, as detailed in section 4.3.2a.

Rv0950c Primers

FW 5' CATATGGCAGCGATTCGCACACCTCG
RV 5' CTCGAGTCAACCGGTGTAATTGCCGACGC

Rv3628 Primers

FW 5' CATATGCAATTCGACGTGACCATCG
RV 5' CTCGAGTCAGTGTGTACCGGCCTTGAAGC

Rv3836 Primers

FW 5' CATATGACAGTACGGATGGACCCGCA
RV 5' CTCGAGTCATGGGCCGTTCATAGCATCGG

Figure 4.16: Primer sequences used to amplify Rv0950c, Rv3628, and Rv3836 for *in vivo* expression. Bold text identifies *Nde*I (red) and *Xho*I (blue) restriction sites. Start/stop codons are underlined and normal text identifies target-unique sequences.

4.3.4b Cloning of target genes

PCR products were first inserted into the cloning vector, pGemT1 (Promega) by TA cloning. Purified PCR product was included, at a vector to insert ratio of 1 : 1, with 50 ng of pGemT1 DNA, 3 U of T4 DNA ligase, and 1 x ligation buffer (all Promega) in a 10 µl reaction. Reactions were incubated at room temperature for 1 hour (Rv0950c and Rv3628) and at 4 °C overnight (Rv3836). 20 µl of Novablue cells (Novagen) were transformed with 1 µl of ligation reaction, as described in section 4.3.2b. Cells were plated onto LB-Amp plates containing 0.5 mM IPTG and 80 µg/ml X-gal, to enable blue/white colony screening.

For each target, 8 single white colonies were grown overnight in LB-Amp and the plasmids were purified as described in section 4.3.2b. Purified plasmids were digested with *Nde*I (New England Biolabs) and *Xho*I (Roche), in preparation for subcloning into the expression vector, pET28a (Novagen). 7.5 µl of plasmid (from 150 µl) was added to a 20 µl reaction with 20 U each of *Nde*I and *Xho*I and 1 x reaction buffer #2 (New England Biolabs). Reactions were incubated at 37 °C for 3 hours and inserts of correct size were identified by agarose gel electrophoresis, as described in section 4.3.2b.

Two positive clones were selected for each target. Both constructs and pET28a plasmid DNA were digested with *Nde*I and *Xho*I, and purified from an agarose gel. 100 ng of digested and purified pET28a was included in a 20 µl reaction with 0.2 pmol of purified insert, 1 U of T4 DNA ligase, and 1 x ligase buffer (both Roche) and incubated overnight at 15 °C. HMS174 (DE3) and Rosetta 2 (DE3) cells were transformed with ligation reaction, as described in section 4.3.2b. Purified plasmids were obtained from overnight cultures from a single colony and two clones for each target were sent to Oxford University for DNA sequencing (http://polaris.bioch.ox.ac.uk/dnaseq). Data were analysed as outlined in section 4.3.2b. One positive clone with the correct sequence for each target was selected and progressed into expression trials.

4.3.4c Expression trials of target genes

Starter cultures, LB-Kan (containing 35 µg/ml kanamycin) for HMS174 (DE3) cells and LB-Kan-Cam for Rosetta 2 (DE3) cells, were prepared as described in section 4.3.2c. Overnight culture was diluted ten fold in fresh TB, containing the appropriate antibiotic, and grown to an OD 600 of 0.6 – 1.0. Cultures were induced with 1 mM of IPTG before incubating at the specified temperature (**table 4.11**), whilst shaking at 185 rpm. The OD 600 was recorded before pelleting cells by centrifugation at 8, 000 rpm in a Beckman JA20 rotor for 20 minutes and storing at –20 °C overnight.

Thawed cells were resuspended in Na-Lysis buffer and disrupted by sonication on ice, using a Branson sonicator at 50 % output. Cells were subjected to twenty cycles of a 3 second burst followed by a 7 second rest, with a resting period at the midway point to prevent overheating of the sample. Soluble fractions were obtained by centrifugation at 14, 000 rpm for 30 minutes and prepared for SDS-PAGE, as described in section 4.3.2c. Expression conditions are described in **table 4.11**.

***E. coli* strain**	**Rationale**	**Time (hours)**	**Temp. (°C)**	**Volume (ml)**	**Lysis method**
HMS174 (DE3)	Initial conditions	6	25 30	50	None
Rosetta 2 (DE3)	Cells may provide unusual codons required for target protein expression/solubility	6	25 30	50	Sonication
Rosetta 2 (DE3)	As above. Increased incubation period may increase expression	24	25	1000	Sonication

Table 4.11: Optimisation of *in vivo* expression conditions for Rv0950c, Rv3628, and Rv3836 in pET28a. Cultures grown in terrific broth (TB).

4.3.4d Large scale expression and purification of Rv3628

The expression of Rv3628 was scaled-up to produce a sufficient quantity of protein for purification by chromatography. The resulting yield was compared with that produced from cell-free expression (section 4.2.3d).

Rv3628 was expressed for 6 hours at 25 °C in Rosetta 2 (DE3). 1L of culture was pelleted and resuspended in 7.5 ml of ppa-Lysis buffer and the soluble fraction was obtained by the French pressure cell method and centrifugation (see section 4.3.2d). The soluble fraction

(10 ml) was applied to a 1 ml HisTrap HP Nickel column (Amersham), equilibrated in ppa-NiA buffer. Non-bound material was washed in 20 column volumes of ppa-NiB buffer at 1 ml/min with a maximum pressure of 0.3 MPa. Bound protein was eluted with 10 column volumes of ppa-NiB buffer and collected in 1 ml fractions. A large peak in 280 nm absorbance was observed after washing with 5 column volumes of eluant. Five fractions were pooled (5 ml) and loaded onto a 5 ml HiTrap Desalting column (Amersham) equilibrated in ppa-AxA buffer. The sample was applied to the column at 2 ml/min with a pressure limit of 0.3 MPa. Finally, the single large peak was pooled (5 ml) and applied to a 5ml HiTrap QFF anion exchange column (Amersham), equilibrated in the same buffer. The column was washed with 5 column volumes of buffer at 5 ml/min, with a maximum pressure of 0.5 MPa. Bound protein was eluted in a gradient of 0 – 100 % buffer ppa-AxB (0 – 1 M sodium chloride) over 15 column volumes. A large, well defined peak in absorbance at 280 nm was observed at 75 % of eluant, was pooled (17 ml) and concentration was determined by Bradford assay.

4.3.5 Cloning and expression of Rv0950c, Rv3628, and Rv3836: Results

All targets were successfully amplified by PCR and positive clones were isolated. Rv3628 was expressed in both Rosetta 2 (DE3) and HMS174 (DE3) cells, however total yield was dramatically reduced when using the latter (**figure 4.17A**). Rv3628 was predominantly expressed in a soluble form by Rosetta 2 (DE3) cells. Extended incubations of the expression media marginally improved solubility of this target. Chromatographic steps removed ~ 95 % of contaminating proteins from the large-scale preparation, yielding 20 mg of pure Rv3628 per litre of culture (**figure 4.18**). Rv0950c (**figure 4.17C**), and to a lesser extent, Rv3836 (**figure 4.17B**) were also expressed in Rosetta 2 (DE3) cultures, however Rv0950c remained insoluble. Solubility of Rv3836 was improved by increasing the incubation time to 24 hours.

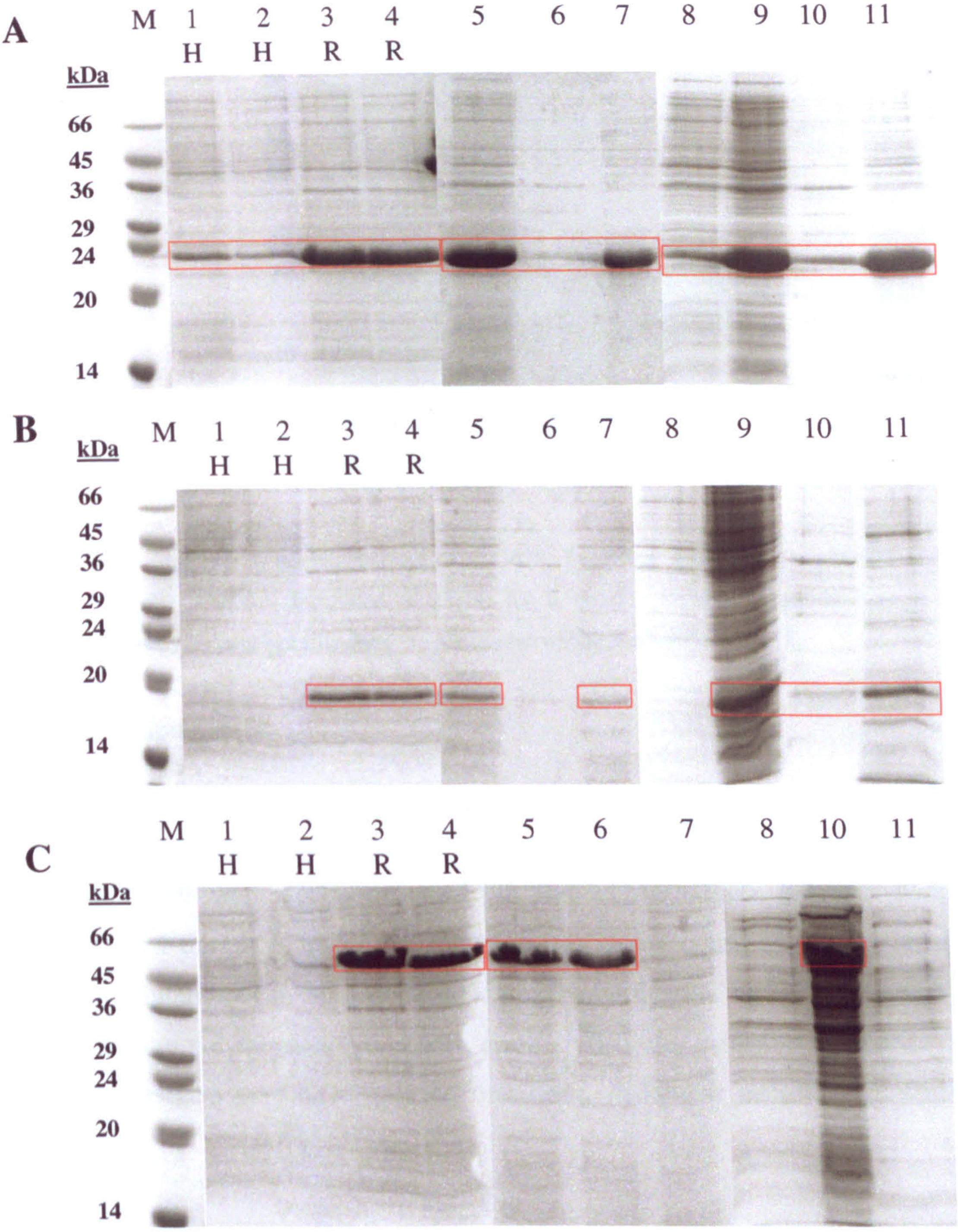


Figure 4.17: 15 % SDS-PAGE from small-scale *in vivo* expression trials of: **(A)** Rv3628; **(B)** Rv3836; and **(C)** Rv0950c, in pET28a (section 4.3.4c). Total extracts from HMS174 (DE3) (H) and Rosetta 2 (DE3) (R) cultures incubated for 6 hours at: **25 °C**, lanes 1 & 3; **30 °C,** 2 & 4. Rosetta 2 (DE3) cultures incubated at 25 °C for: **6 hours** total (lane 5), insoluble (6), soluble (7); **24 hours** uninduced (8), total (9), insoluble (10), soluble (11). Molecular weight markers in lanes M. Target proteins highlighted in red.

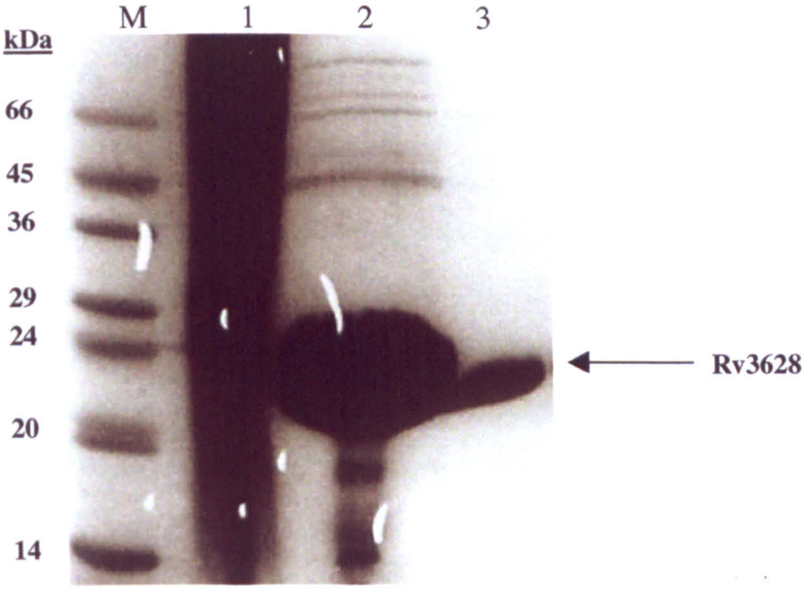


Figure 4.18: Fractions from purification of Rv3628, following large-scale production and purification (section 4.3.4d). Cultures incubated at 25 °C for 6 hours. **Soluble lysate** in lane 1; **post-affinity chromatography**, 2; **post-gel filtration**, 3. Molecular weight markers in lanes M.

4.4 Discussion

4.4.1 Summary of results from the cell-free system

Positive clones were obtained for all but five of the targets, however with additional experimental time it is likely that mutation-free clones could have been obtained for all targets. Mutations can be introduced during amplification by PCR, although the use of proof-reading polymerases limits such errors. Also, the inclusion of DMSO during PCR for some targets may have induced mutations, making it necessary to sequence more clones.

The cell-free system proved to be a highly successful method for rapidly screening large numbers of *M. tb* targets for expression. Initial screening, using standard reaction conditions without additives, resulted in one third of the targets yielding soluble protein. For targets where expression was not observed and discrete PCR products were not available, inclusion of DMSO during PCR resulted in the correct amplification of PCR 2 templates, which subsequently led to the production of soluble protein for a further three targets.

Metals are known to act as structural elements and folding nucleation points for unfolded proteins (ITQB/UNL, 2006). Due to the predicted metal requirement for several of the targets, it was possible that the lack of such ions within the reaction solution might account for insolubility. However addition of metal compounds to the cell-free reactions did nothing to improve solubility, with the exception of a marginal improvement in solubility for Rv2845c upon addition of zinc. A number of the targets have unknown metal requirements and so it may be necessary to screen a larger number of different ions. Also, optimisation of ion concentration and the inclusion of proteins involved in metal incorporation may be required to induce any noticeable improvement in the expression of soluble protein.

Addition of detergents or molecular chaperones to the standard reaction conditions greatly improved solubility for problematic targets. Five previously insoluble targets were made soluble by the additon of detergents, and a further four by molecular chaperones, bringing the total percentage of targets expressed in a soluble form to 67 %. Considering the equivalent success rate for the TBSGC is 30 % (soluble expression obtained from 518 out of 1558 targets, as of December 2006), this constitutes a significant improvement. Little difference was observed between the two detergents tested, however the dnaJ-dnaK-grpE group of molecular chaperones proved to be the most successful in producing soluble *M. tb* proteins (through promotion of correct folding), followed by the groEL-groES pair. Taking into account the costs associated with these additives (the preparation of chaperones for use in cell-free reactions is time consuming and detergents such as Digitonin are extremely expensive) and potential difficulties during their removal, synthesis by these methods may only be cost-effective when significant yields are obtained.

Milligram quantities of soluble protein were obtained for 9 of the 28 original targets screened (section 4.2.3), however only two of the newly selected 8 targets gave similar results (section 4.2.5). Although some soluble expression was observed for all of the previously insoluble targets, by addition of detergents or molecular chaperones to the reaction solutions (section 4.2.5c), only Rv2388c was significantly improved. One target, Rv0950c, was thought to be soluble throughout expression trials. However, upon purification by affinity chromatography, the protein was 'lost' suggesting either degradation during the purification stage or a misinterpretation of solubility during SDS-PAGE analysis.

4.4.2 Summary of results from the *in vivo* system

Several targets were selected for comparative expression trials using a traditional *in vivo* system, as described in section 4.3.1. Targets were cloned into expression vectors which were used to transform two expression hosts, HMS174 (DE3) and Rosetta 2 (DE3). Positive clones were obtained for all targets. Whilst both Rv3628 and Rv3545c were expressed in HMS174 (DE3), yield was significantly improved by using Rosetta 2 (DE3) hosts, and expression of Rv3836 and Rv0950c was only achieved using these cells. This is likely to be due to Rosetta 2 cells providing tRNA molecules with the unusual codons necessary to translate *Mycobacterium tuberculosis* proteins efficiently (Del Tito *et al.*, 1995, Rosenburg, 1996, and Novy *et al.*, 2001). No difference in total expression was observed between cultures incubated at 25 °C and those at 30 °C, but it seems sensible to use lower temperatures in order to facilitate solubility (Novagen, 2006). Whilst Rv3836 was expressed in a soluble form, overall yield was very low, and Rv0950c remained insoluble throughout.

Rv3628 expressed very well without optimisation of the expression conditions. However only minimal amounts of soluble Rv3545c were produced and altering the expression temperatures (18 – 30 °C), incubation times (6 - 24 hours), and host cell (HMS174 and Rosetta 2) made little difference to the overall solubility of this target. A literature search of homologous P450s (such as Bellamine *et al.*, 1999 and McLean and Cheesman *et al.*, 2002) found that soluble protein was generally obtained 6 to 24 hours after induction, but soluble yields of Rv3545c were extremely low under these conditions and were only improved by increasing the incubation to 72 hours. Such extended expression periods are only possible in nutrient-rich media such as terrific broth which contains both glycerol and a high proportion of yeast extract. The effect of incubation time on Rv3545c expression is clearly demonstrated in **figure 4.15A**, which shows cell pellets after 24 (light brown) and 72 hours (dark red/brown).

4.4.3 Advantages and disadvantages of the cell-free expression system

One of the major benefits of this system is the ability to express target proteins either directly from the PCR product or from constructs generated in just one cloning step (Lesley *et al.*, 1991). This significantly reduces the amount of time taken to progress from target selection to protein expression. Tags to aid solubility or purification are easily incorporated by tailoring PCR primers to the specific requirement. The small volumes of reaction solution required to produce quantities of protein sufficient for analysis by SDS-PAGE also allow for the initial screening of large numbers of targets in a short time frame.

Problems associated with the expression of proteins toxic to host cells is not an issue with this system, due to the lack of whole cells (Golf and Goldberg, 1987 and Murthy *et al.*, 2004). Particularly important when attempting to express proteins from species such as *M. tb,* is the ability to translate sequences which utilise non-*E. coli* codons, and this is another advantage of the system, due to the optional inclusion of minor-tRNA's within the reaction solution (Chumpolkulwong *et al.*, 2006). Additional benefits are the ease with which additives and cofactors can be included in the reaction solutions and subsequently screened against target proteins for solubility/correct folding (Betton, 2003 and Murthy *et al.*, 2004). Proteins can also be synthesised in a methionine-deficient environment, thus ensuring full selenomethionine incorporation when MAD phasing is required for structure determination (Kigawa *et al.*, 2002).

Whilst the cell-free system proved to be a highly successful method for expressing *M. tb* proteins in this study, there are also disadvantages. The system is generally regarded as being more expensive to set up than the *in vivo* method, although the success rate may outweigh this cost. High grade reagents and cell extract must be either commercially obtained or produced in-house, the latter of which is a time consuming and technically challenging process (Murthy *et al.*, 2004). The success achieved during this project was facilitated by performing the experiments in a laboratory dedicated to *in vitro* expression systems. Finally, the expression of certain metalloproteins, which are biologically active, using this system may not be possible due to a lack of metal incorporation pathways. However, many metalloproteins have been successfully expressed using the cell-free system, and a recent publication describes an optimised protocol for the expression of zinc-binding proteins (Matsuda *et al.*, 2006).

The cell-free system operates a more high-throughput approach than most traditional systems and its potential to screen large numbers of targets is one of its major advantages. However, although it is possible to screen many different reaction conditions in an attempt to improve negative results, larger starting sample pools clearly result in a greater success. Moreover, it seems that targets that express without assistance are more likely to produce satisfactory quantities of protein when synthesised on a large scale. That is not to say it is impossible to produce sufficient amounts of these proteins, only that more time would be required to optimise the scaling-up of a smaller number of targets, and this must be balanced against the cost of such procedures.

4.4.4 Advantages and disadvantages of the *E. coli in vivo* expression system

The key benefit of this system is the ease with which it can be set up and executed in virtually any biological laboratory. Standard protocols are well established and the necessary reagents are readily available from commercial sources. Additional buffers and growth media can be made in-house using standard laboratory chemicals and are easily disinfected by autoclaving at high temperatures. This all accounts for the relatively low cost required to establish such a system.

As this method utilises whole living cells, it is often unnecessary to include additives such as metals which are required for solubility/folding or activity, as these are available within the system. However, when overexpressing metalloproteins the requirement can exceed the supply and the additive or a precursor may be included for optimal results. Once positive clones have been obtained and a successful expression strategy has been optimised, expression can easily be scaled up using glycerol stocks of the original colony and repeated whenever necessary to obtain more protein.

A disadvantage associated with the *in vivo* system can be the length of time taken to progress from gene to protein (Murthy *et al.*, 2004). More than one cloning step is generally required, although PCR products can be directly cloned into pET expression vectors (Novagen, 2006). Also, it may be necessary to screen a number of vector-host combinations to produce target protein. This can make the high-throughput expression of targets extremely time consuming and whilst the procurement of reagents and equipment may be of minimal cost, such experiments can be financially unviable due to the intensive labour required. *In vivo* methods are however regularly used for high-throughput protein production, due to the availability of efficient plasmids, which improve solubility and simplify purification, and highly inducible gene expression hosts (Braun and LaBaer, 2003, Scheich *et al.*, 2003, and Murthy *et al.*, 2004).

As described above, an extensive but often expensive selection of host cells are commercially available, optimised for particular stages of cloning and expression. Alternatively stocks of host cells can be grown in-house and made competent for expression, but the convenience of ready-made cells may be preferred. This comprehensive range of cells are generated either by modifications to the host's genome or by transforming the bacteria with plasmid DNA, both of which provide mechanisms to aid the soluble expression of target proteins. Such modifications may provide the unusual codons required to translate a target protein or promote correct folding by enhancing disulphide bond formation within the cytoplasm (Novagen, 2004). An even wider selection of bacterial plasmids are commercially available which offer a multitude of cloning options. Target genes are conveniently inserted into the multiple cloning site of such plasmids and proteins can be expressed, together with tags required for signalling or purification, however these can also be added by designing suitable primers during PCR.

Whilst the scaling up of expression conditions is in principle very simple, many litres of culture may be required to yield significant quantities of target protein and not all laboratories are equipped with such facilities. Target protein may be retained in the cytoplasm, exported to the periplasmic space, or secreted into the medium, and extraction can form a bottleneck. Mechanical disruption by a French pressure cell can be difficult to operate, time consuming, and not suitable for high-throughput purposes. Also, chemical lysis methods are not always successful. Finally, purification of target proteins from the substantial *E. coli* contribution can be difficult, particularly when the overexpressed yield is low.

4.4.5 Comparisons between the cell-free and *in vivo* expression systems

Whilst Rv3836 was expressed in a soluble form using the *in vivo* system described, overall yield was very low. As with the insoluble Rv0950c, further optimisation of the expression conditions may have improved this. However, the similar results obtained using the cell-free system suggest that improvements may not be possible.

Twenty milligrams of pure protein per litre of culture was obtained for Rv3628 using the *in vivo* system. Comparing this with the total yield from cell-free synthesis of 27 mg (1.5 mg per millilitre of reaction solution), the *in vivo* system does appear to have the advantage in that the cost and effort required to scale up such a system would be minimal. Another issue to raise during this comparison is that of metal incorporation. Rv3628, an inorganic pyrophosphatase, requires metals such as magnesium or zinc for enzymatic activity (Cooperman and Chiu, 1973 and Lahti and Kolakowski *et al.*, 1990). The X-ray crystal structure from protein synthesised using the cell-free method (see chapter 5), reveals the lack of these metals and so represents an inactive molecule. Magnesium acetate (9.28 mM) was included in the synthesis reaction, however this was clearly not incorporated into the protein. Greater quantities of metal ions may have been required, however due to the essential requirement of magnesium for RNA polymerase activity, levels of these ions were optimised for expression and so alteration may affect overall productivity. Another option would be to incorporate metals during crystallisation or to soak the apoenzyme crystals in metal solutions prior to data collection. These methods were used to prepare Y-PPase (Heikinheimo *et al.*, 1996 and Harutyunyan *et al.*, 1996) and E-PPase (Samygina *et al.*, 2001) holoenzymes, prior to X-ray data collection, suggesting activating metals were not fully incorporated during protein expression.

The greatest success from this comparative study was the expression of Rv3545c, a cytochrome P450 125, which yielded 13 mg of pure protein per litre of culture, and yet was predominantly insoluble using the cell-free system. The ability of cell-free systems to express active haem proteins is questionable, as there are no mechanisms whereby haem can be produced or incorporated into the polypeptide. In living cell systems, a precursor is converted to haem by its biosynthetic pathway, which is not possible in *in vitro* set ups due to the lack of whole cells.

To summarise, the cell-free system was used to rapidly express large numbers of targets, however the *in vivo* system was arguably the more successful. Of the four targets chosen, three gave comparable results to those obtained from cell-free synthesis, and the expression of Rv3545c was dramatically improved (**table 4.12**). However, a clear benefit of the cell-free system is shown in section 4.2.5c, whereby relatively simple optimisation of small-scale reaction conditions yielded partially soluble protein for all of the 9 targets tested. Similar results may be obtainable in an *in vivo* system, although optimisation of expression conditions for multiple targets could be time-consuming (Murthy *et al.*, 2004).

		Cell-free expression system		***In vivo* expression system**	
Target	**Function**	**Result**	**Yield**[1]	**Result**[2]	**Yield**[3]
Rv0950c	Probable metalloprotease	Possibly soluble (aggregated during purification?)	n/a	Insoluble	n/a
Rv3545c	Cytochrome P450 125	Soluble (with molecular chaperones)	n/a	Very soluble (72 hour expression)	80 mg (13 mg/L)
Rv3628	Inorganic pyrophosphatase	Very soluble (without additives)	26.9 mg (1.5 mg/ml)	Very soluble (6 hour expression)	20 mg (20 mg/L)
Rv3836	Zinc metalloprotease	Soluble (without additives)	6.0 mg (0.3 mg/ml)	Soluble (24 hour expression)	n/a

Table 4.12: Comparison between cell free and *in vivo* systems for the expression of four *Mycobacterium tuberculosis* targets. [1]Total yield from large-scale preparation. Text in parenthesis shows the yield per millilitre of reaction solution. [2]All soluble proteins expressed in Rosetta 2 (DE3) cells. [3]Total yield from large-scale preparation. Text in parenthesis shows the yield per litre of culture.

Due to the lack of whole cells, the cell-free system would suggest fewer host proteins are present, which in principle would allow for an easier purification, however this was not apparent during this study. A similar number of purification steps were required to remove ~ 95 % of host contaminants, following cell-free expression, as would generally be used following *in vivo* expression. An example is shown in **figures 4.7** and **4.18**, whereby Rv3628 was purified by 3-step chromatography following cell-free expression, and by only 2-steps following *in vivo* expression. However the anion exchange step was routinely performed after cell-free expression and may not have been necessary. The correct engineering of tags which aid purification (such as His-tags) during cloning stages and the use of efficient purification resins can be successful in removing the majority of host contaminants in a single step, as shown in **figure 4.14B**. Furthermore, a study comparing the expression of 63 proteins from *Pseudomonas aeruginosa* using *E. coli*-based *in vivo* and *in vitro* systems, identified a higher proportion of contaminating host proteins for *in vitro*-expressed proteins, following a single-step affinity purification (Murthy *et al.*, 2004).

4.4.6 Additional considerations

Whilst not an inherent issue with the cell-free system, another problem encountered during this study was the inability to replenish protein stocks once experimental time at RIKEN had ended. This was due to the lack of specialist equipment and reagents at Daresbury Laboratory, required to perform such experiments. This resulted in the use of protein, which had been stored for several months, in downstream applications such as crystallisation and enzymatic assays. Also shipping the protein from Japan to England took considerable periods of time, further increasing the chances of degradation and precipitation.

Chapter 5 - Structure of inorganic pyrophosphatase (Rv3628) from *Mycobacterium tuberculosis*

5.1 Introduction

Inorganic pyrophosphatases (PPases) belong to the phosphatase superfamily and are ubiquitous enzymes, which play an essential role in biosynthetic reactions. PPases catalyse the hydrolysis of pyrophosphate (PPi), a product of reactions such as protein and nucleotide synthesis, to orthophosphate $(Pi)_2$. The first PPase X-ray crystal structure was published in 1981 by Arutiunian *et al* (PDB code: 1PYP). This 3.0 Å yeast apoenzyme identified the now distinctive oligomeric fold and large polar active site, characteristic of type I PPases. It was a further 13 years before the second structure became available, a high resolution 2.0 Å structure of *Thermus thermophilus* PPase (PDB ID: 2PRD).

Much of our existing knowledge of this enzyme has been gained in the last few years, 34 apoenzyme/ligand-bound structures are now deposited in the Protein Data Bank (PDB), from nine unique organisms: *Escherichia coli* (1IPW & 1FAJ: Kankare, 1996; 1OBW: Harutyunyan, 1997; 1JFD & 1MJW: Avaeva, 1997 & 1998; 1I40 & 1I6T: Samygina, 2001;), *Saccharomyces cerevisiae* (1WGI/1WGJ & 1E6A/1E9G: Heikinheimo, 1996 & 2001; 1YPP: Harutyunyan, 1997; 8PRK: Tuominen, 1998; 1M38: Kuranova, 2003), *Bacillus subtilis* (1K23: Ahn, 2001; 1WPM/1WPN: Fabrichniy, 2004), *Streptococcus gordonii* (1K20, Ahn, 2001; 1WPP: Fabrichniy, 2004), *Streptococcus mutans* (1I74: Merckel, 2001), *Sulfolobus acidocaldarius* (1QEZ: Leppanen, 1999), *Pyrococcus furiosus* (1TWL: Zhou, 2004), *Pyrococcus horikoshii* (1UDE: Liu, 2004), and *Thermus thermophilus* (2PRD: Teplyakov, 1994). 19 of these 34 structures were solved in the last 7 years, predominantly due to advances in genomic sequencing.

A detailed introduction to inorganic pyrophosphatases is given in section 1.2. The X-ray crystal structure of Mtb-PPase is presented here, refined to a resolution of 2.7 Å. Analysis of the structure and comparisons with a number of the existing structures mentioned above, are presented here.

5.2 Protein expression and crystallisation

Mtb-PPase was synthesised using the cell-free expression method outlined in section 4.2.2g. Crystals were grown at room temperature using the hanging drop vapour diffusion method. The protein concentration was 5 mg/ml in CF-E buffer (50 mM NaH_2PO_4 pH 8.0, 150 mM NaCl, and 1 mM DTT). 2 μl of protein was mixed with an equal volume of reservoir solution containing 2 M ammonium sulphate, 2 % v/v PEG 400, and 0.1 M HEPES-Na, pH 7.5. Crystals grew in one week to a size of 100 x 200 μM. The multiple crystal shown in **figure 5.1** was separated using a thin brass needle, and the larger of the two crystals was used to collect the full dataset. Separation and mounting of the crystal were performed by Dr Svetlana Antonyuk at the SRS.

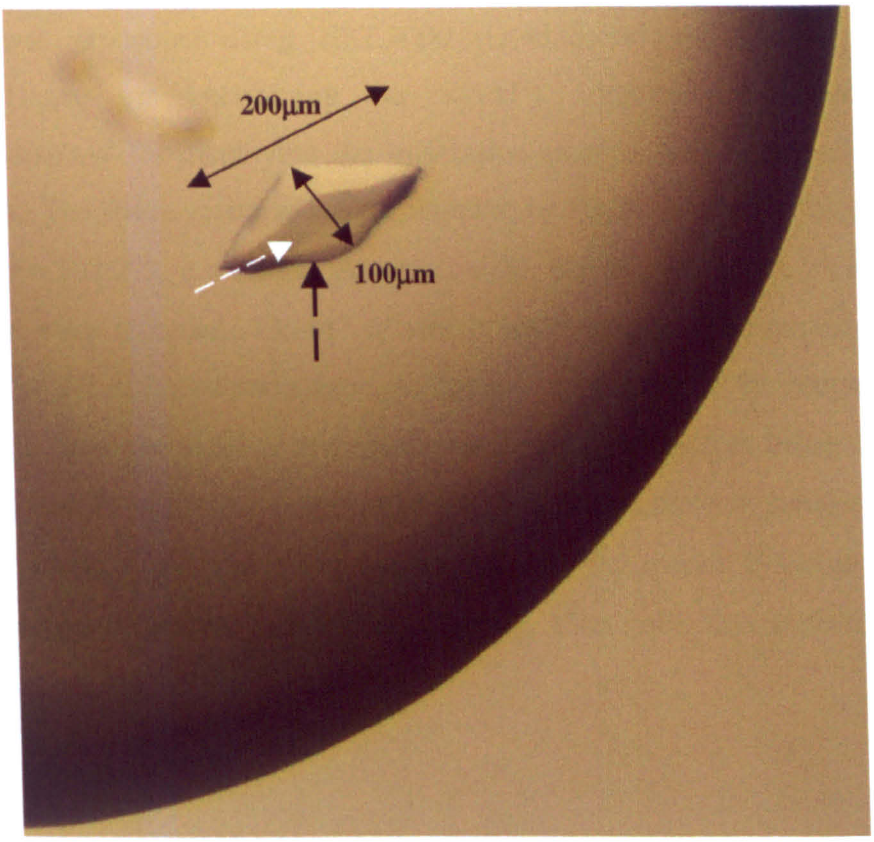


Figure 5.1: The multiple Mtb-PPase crystal, which was separated and used during data collection. The larger of the two crystals (indicated with a broken black arrow) was separated from the smaller crystal (white broken arrow) and used to collect data.

5.3 Data collection

Prior to data collection the crystal was soaked for approximately one minute in a cryoprotectant solution consisting of mother liquor and 25 % glycerol, before flash-cooling to 100K in a nitrogen cryostream. X-ray diffraction data were collected on station MAD 10.1 at the Synchrotron Radiation Source, Daresbury (Cianci *et al.*, 2005) using a Mar CCD detector at an X-ray wavelength of 0.979 Å. The maximum resolution obtainable was 2.7 Å. The crystal to detector distance was set to 265 mm and data were collected over an oscillation angle of 1 ° per image, with an exposure time of 60 seconds per frame. A total of 360 images were collected. See **figure 5.2** for the first diffraction image.

5.4 Data processing

The dataset was processed using HKL2000 (Otwinowski and Minor, 1997), which incorporates Denzo for determining the crystal's orientation and performing spot integration, Xdisplayf for displaying the diffraction images, and Scalepack to scale and merge the data. The space group was determined to be $P3_221$, with unit cell parameters of: a = 102.022, b = 102.022, c = 80.812 Å; and $\alpha = 90$, $\beta = 90$, $\gamma = 120$ °. A total of 1, 234, 197 reflections were recorded, 13, 447 of which were unique. The maximum resolution was judged to be 2.7 Å, based upon three parameters: data being 98 % complete, with 87 % completeness in the outer shell at this resolution; merging R-factors being 16.1 % for the whole dataset, and 63 % for the outermost shell; and an intensity to standard deviation of intensity ratio, $I/\sigma(I)$, of 6.1 and 1.5 in the outer shell. The overall B-factor was estimated using a Wilson plot to be 32.94 $Å^2$ (Wilson, 1949). Data collection statistics are given in **table 5.1**

Figure 5.2: The first diffraction image from the Mtb-PPase crystal. **(A)** The whole image, with labels representing the resolution shells to 50, 10, 5, and 2.5 Å. **(B)** A region near to the resolution limit of 2.7 Å. Image generated using HKL2000 (Otwinowski and Minor, 1997).

Resolution range (Å)	44.00 - 2.70
Space group	P3$_2$21
Unit cell parameters	a = 102.022, b = 102.022, c = 80.812 Å $\alpha = 90$, $\beta = 90$, $\gamma = 120$ °
Redundancy (last shell)	5.6 (3.2)
Average *I/σ(I)* (last shell)	6.1 (1.5)
Rmerge (%) (last shell)	16.1 (63.0)
Completeness (%) (last shell)	98.0 (87.0)
Reflections	
(Overall)	1, 234, 197
(Unique)	13, 447
Wilson B value (Å^2)	32.94

Table 5.1: Mtb-PPase data processing statistics.

5.5 Solvent content

The contents of the asymmetric unit were estimated using the Matthews coefficient calculation (**equation 5.1**). The molecular weight of monomeric Mtb-PPase is known to be 18.3 kDa.

Equation 5.1:

$$V_m = \frac{V_{cell}}{NM_r}$$

Where V_{cell} is the volume of the unit cell, N is the number of monomers in the asymmetric unit, and M_r is the molecular weight of the protein. The Matthews coefficient was calculated to be 2.2 Å^3 Da^1. If three monomers are located in the Mtb-PPase asymmetric unit, the associated solvent content is 50 %.

5.6 Structure solution and refinement

The structure was solved by the molecular replacement method using MOLREP (Vagin and Teplyakov, 1997), as part of the CCP4 suite (CCP4, 1994). The 2.2 Å crystal structure of a monomeric *E. coli* pyrophosphatase (PDB ID: 1FAJ) (Kankare *et al.*, 1996), which shares

45 % sequence identity with Mtb-PPase, was used as the starting model. Initially, the most plausible solution for only one monomer was determined. This was then inputted as a fixed solution during determination of the remaining two monomers within the asymmetric unit. This yielded an overall R-factor of 51 % with a correlation coefficient of 39.8 %.

The model was rebuilt using COOT (Emsley and Cowtan, 2004) and O (Jones *et al.*, 1991) and refinement was performed using the maximum likelihood method in REFMAC5 (Murshudov *et al.*, 1997), both using an iMAC G5 machine (Apple). Throughout refinement PROCHECK (Laskowski *et al.*, 1993) and WHATIF (Vriend, 1990) were used to check the model's stereochemistry.

Initial rigid body refinement, using data in the range of 44 to 3 Å, slightly reduced the R-factor to 50 % (R-free 50 %) after twenty cycles. Prior to this, 5 % of the reflections were set aside for determination of the free-R factor (Brünger, 1992). Medium main chain and loose side-chain non-crystallographic symmetry restraints were applied to the three monomers in the asymmetric unit, during restrained-positional and individual isotropic temperature refinement. With the weighting matrix set to 0.01, the R-factor was further lowered to 36 % (R-free 44 %) after 20 cycles. Additional data to the maximum resolution of 2.7 Å was then subjected to another 20 cycles of refinement, with the weighting term increased to 0.02 and 0.03, resulting in R-factors of 34 and 33 % respectively (R-free 44 and 45 %). Successive cycles of refinement and examination of the model's fit within the electron density, with the weighting term set to 0.05, decreased the R factor to 28 % (R-free 38 %).

The ARP/wARP program within CCP4 (Lamzin, 1993) was then used to identify potential solvent peaks. Three cycles of ARP/wARP refinement followed ten restrained cycles, resulting in the addition of 42 waters and an R-factor of 23 % (R-free 33 %). The modelling of a further 13 water molecules gave an R-factor of 22 % (R-free 32 %). This was followed by multiple cycles of maximum likelihood and restrained TLS refinement (Winn *et al.*, 2001) and the removal of unfeasible water molecules, resulting in an R-factor of 21 % (R-free 30 %).

In an attempt to improve the R-free, tight non-crystallographic symmetry restraints were applied to both the main and side chains. Unfeasible water molecules were also removed, resulting in a final R-factor of 23.2 % (R-free 27.4 %). The final refined model included 36 water molecules, with an average B-factor of 19.1 Å^2. Water molecules were only modelled when the following criteria were met: well-defined positive peaks were visible in both the 2Fo-Fc and Fo-Fc density maps; reasonable hydrogen bonds with protein residues or other water molecules; and with acceptable temperature factors in relation to the average solvent B-factor. See **table 5.2** for a summary of the refinement process and **figure 5.3** for a plot of R-factors as a function of refinement cycle.

Cycle number	Maximum resolution (Å)	R-factor (%)	R-free (%)	Solvent molecules	Description
1	3	50	50	0	Rigid body refinement
2	3	36	44	0	Restrained positional and isotropic refinement (weighting term 0.01). Medium main chain/loose side chain NCS restraints
3	2.7	34	44	0	As above (weighting term 0.02)
4	2.7	33	45	0	As above (weighting term 0.03)
5	2.7	29	37	0	Model rebuild and refinement (weighting term 0.03)
6	2.7	28	38	0	As above (weighting term 0.05)
7	2.7	23	33	42	Addition of waters by ARP-wARP refinement
8	2.7	22	32	53	Addition of more waters by ARP-wARP refinement
9	2.7	21	30	44	TLS refinement and removal of unfeasible water molecules
10	2.7	23.2	27.4	36	TLS and tight NCS-restrained refinement. Removal of unfeasible water molecules

Table 5.2: The process used to refine the Mtb-PPase crystal structure.

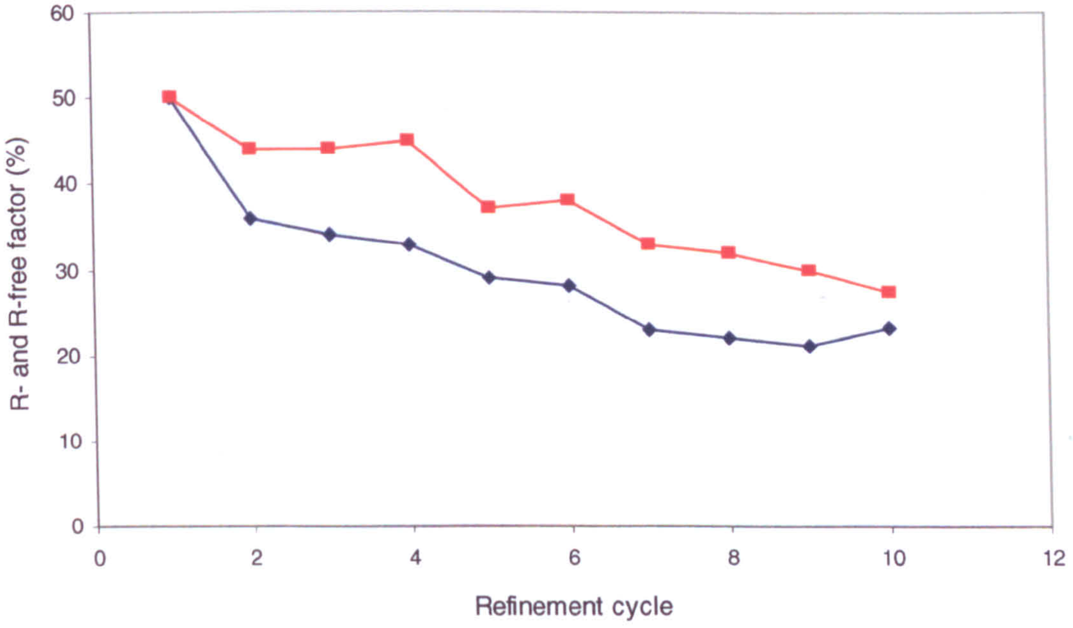


Figure 5.3: R-factors (%) as a function of refinement cycle number. The R-factor is shown in blue and the R-free in red.

5.7 Quality assessment

The stereochemical quality of the model was assessed using PROCHECK (Laskowski *et al.*, 1993) and WHATIF (Vriend, 1990). A Ramachandran plot, generated using CCP4 (Ramachandran *et al.*, 1963), found 90.3 % (363 residues) of all non-proline or glycine residues to be in the "most favoured" regions, with 9.5 % (38 residues) within the "additionally allowed" region. Ala145 (chain B) was found within the "generously allowed" region.

The estimated standard uncertainty (ESU) (Cruickshank, 1996) based upon the R-free value was found to be 0.417 Å. This describes the contribution of experimental data to the positional uncertainty of an atom with a B-factor equal to that of the Wilson B value for the whole molecule.

Refinement and model quality statistics are summarised in **table 5.3** and the Ramachandran plot is shown in **figure 5.4**.

Matthews coefficient (Å^3 Da^1)	2.2
Solvent content (%)	50
Protein atoms	3828
Solvent atoms	36
Rwork (%)	23.2
Rfree (%)	27.4
B value (Å^2)	
Average	19.1
Solvent	32.1
Phosphate	30.6
rmsd bonds (Å)	0.012
rmsd angles (°)	1.362
Ramachandran plot (%)	
Most favoured regions	90.3
Additionally allowed regions	9.5
Generously allowed regions	0.2
Disallowed regions	0.0
E.S.U based upon R-free (Å)	0.417

Table 5.3: Mtb-PPase refinement and model quality statistics.

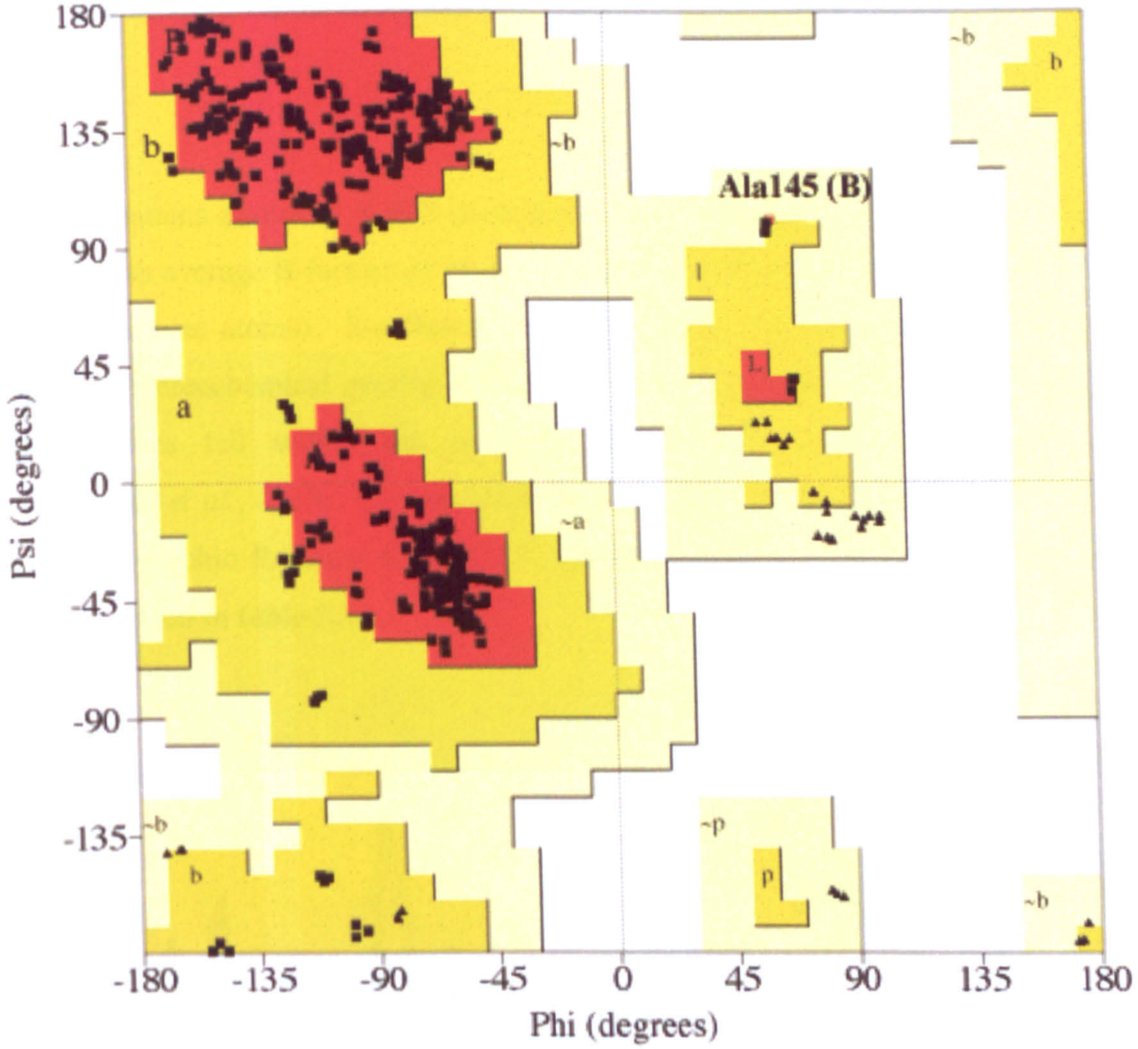


Figure 5.4: The final Ramachandran plot for Mtb-PPase following refinement, generated by PROCHECK using the CCP4 suite (Ramachandran *et al.*, 1963). See text for description.

5.8 Analysis and comparison with other inorganic pyrophosphatases

5.8.1 Quality of the structure

The structure of Mtb-PPase was refined to a maximum resolution of 2.7 Å, consisting of three monomers in the asymmetric unit, each containing one phosphate group within the active site. The final model consists of residues Gly7 to Ala166, from a total of 169 residues in the protein sequence (residues Gly1 to Gly7 form a linker sequence, introduced by the cell-free synthesis method). The remaining residues at the N-terminus were disordered. Side chain density was missing for residues: Lys143 from Cγ onwards (chains B & C); Arg150 from Cζ (C); Arg159 from Cδ (A); Glu162 from Cδ (A & B), and these were modelled with zero occupancy. The side chain density for Phe121 (chain C) was partially missing from Cγ onwards, and these atoms were modelled with an occupancy of 0.5.

The model contains a total of 3, 828 protein atoms, 36 water molecules, and 3 phosphate molecules, with average B-factors of 19.1 Å^2 (overall model), 32.1 Å^2 (solvent atoms), and 30.6 Å^2 (phosphate atoms). See **figure 5.5** for a plot of the average B-factors for each residue. The stereochemical quality of the model was generally good: 90.3 % of all non-glycine residues fell within the most favoured region of a Ramachandran plot (Ramachandran *et al.*, 1963), with no residues in the disallowed region (**figure 5.4**). The final crystallographic R-factor was 23.2 % and the R-free was 27.4 %. Model quality statistics are given in **table 5.3**.

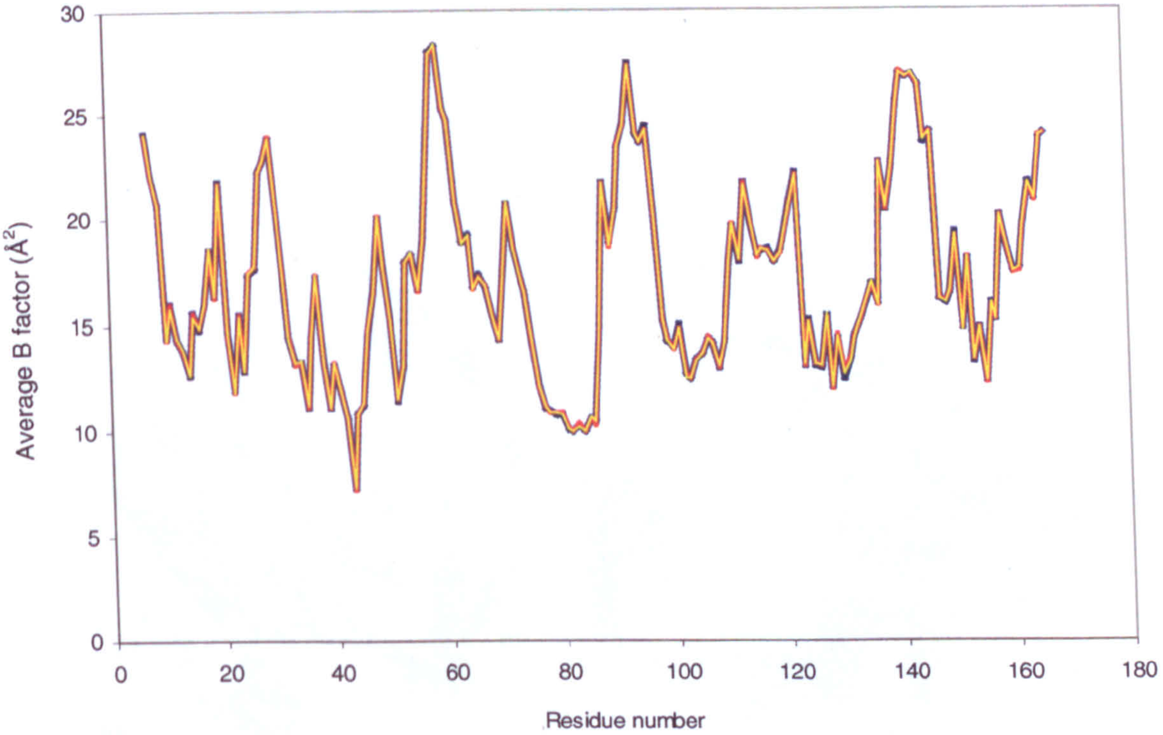


Figure 5.5: A plot of the average amino acid temperature factors for each Mtb-PPase chain: chain A (blue); B (red); and C (yellow). Data generated using the BAVERAGE program in CCP4 (CCP4, 1994). Residues Gly1 to Gly7, which form the linker sequence added during cell-free expression, are not included as they were not visible in the electron density.

5.8.2 Overall structure and oligomeric form

The asymmetric unit of Mtb-PPase consists of three 18.2 kDa monomers, forming a compact non-crystallographic trimer (**figure 5.6**). The monomers are related by a three-fold non-crystallographic symmetry axis and the Cα atoms may be superimposed with root mean square deviations (rmsd) of 0.08 Å.

All crystal structures in this chapter were generated using PYMOL (DeLano Scientific), and labeled secondary structural elements were defined by PROMOTIF (Hutchinson & Thornton, 1996).

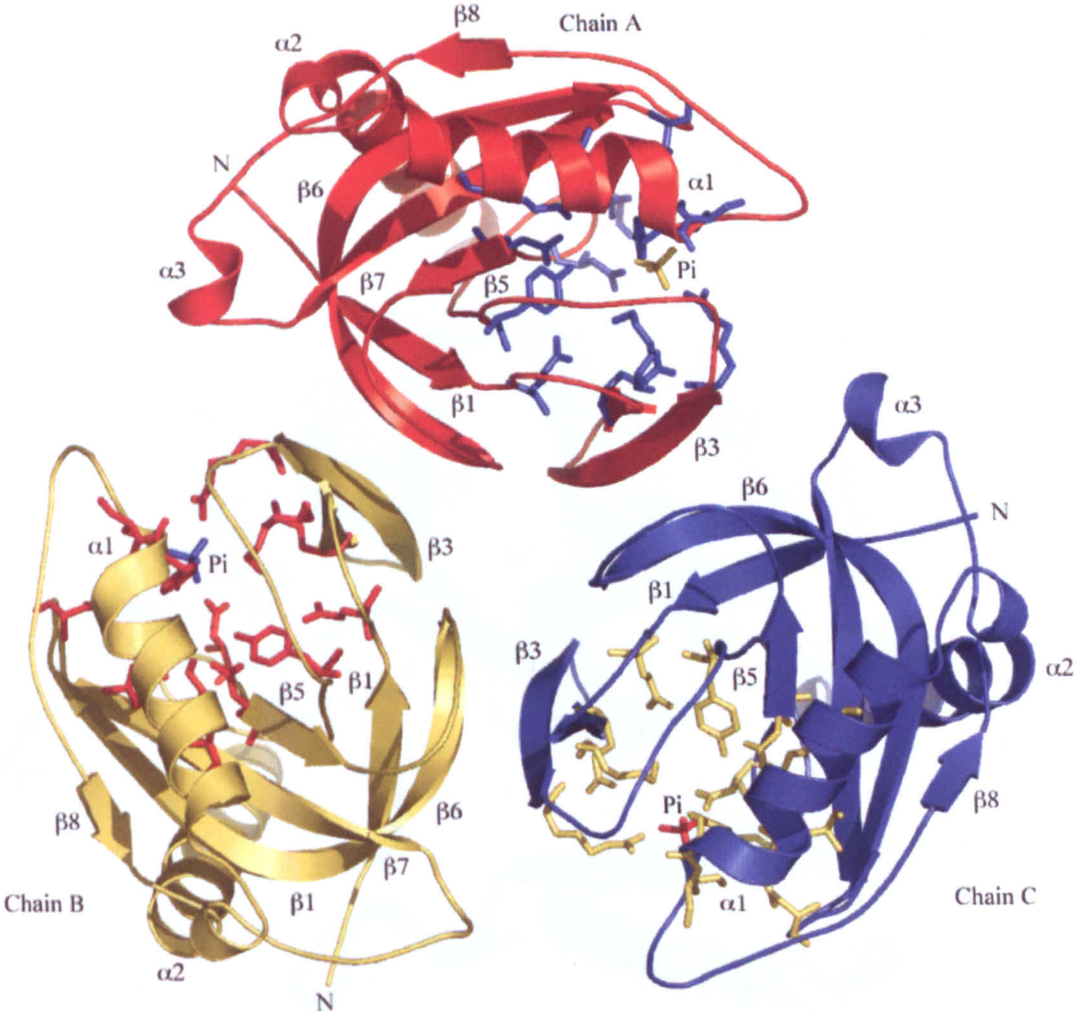


Figure 5.6: Cartoon representation of the non-crystallographic Mtb-PPase trimer. Secondary structural elements are labelled and active site residues are represented as sticks. A nine-strand β-barrel is formed at the centre of the trimer by strands β1, 3, and 6.

The topology of the model was analysed using PROMOTIF (Hutchinson & Thornton, 1996) and is similar to that described for other PPases. The overall fold of each monomer is that of a globular oblate shaped molecule, with each monomer forming a highly distorted β-barrel consisting of strands β1 and β4 to 7; a ten-residue loop connecting β5 and β6; and a 15-residue α-helix (α2) following strand β8, which caps the end of the barrel (**figure 5.7**). A network of hydrogen bonds support the β-barrel. A second 15 residue α-helix (α1), three β-strands (β2 to 3 and β8), and two short α-helices (α3 and 4) surround the β-barrel. Both long helices (α1 and 2) are strictly conserved in all type I PPases, including the much larger 32.2 kDa Y-PPase (Heikinheimo *et al.*, 1996).

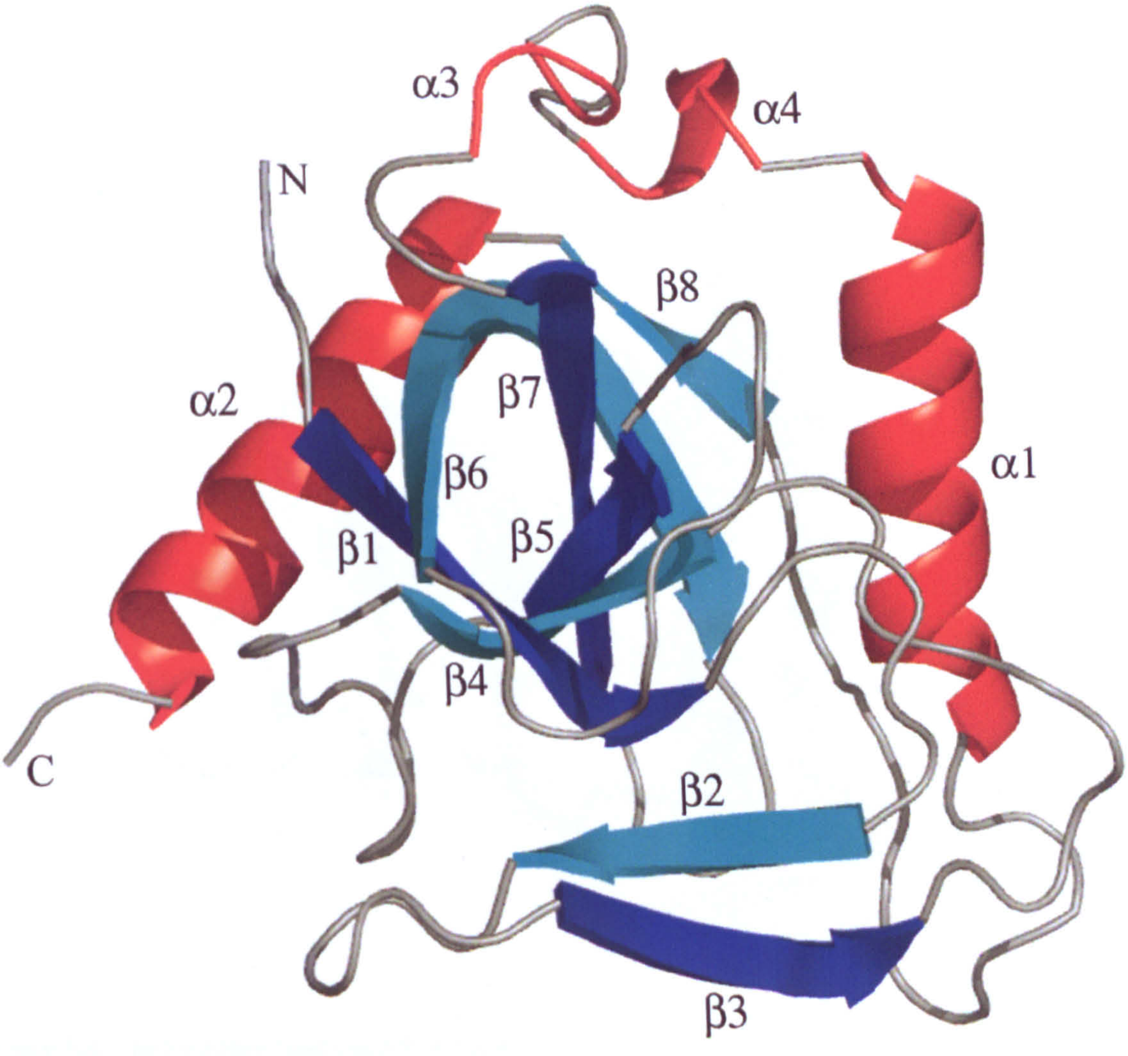


Figure 5.7: A cartoon representation of the Mtb-PPase chain A. Secondary structural elements have been labelled, with β-sheets shown in blue/cyan and α-helices in red. Characteristic of type I PPases, a highly distorted 5 strand β-barrel is formed by strands β1 and β4 to 7.

Each monomer forms a parallel β-sheet with the contiguous monomer, via main chain hydrogen bonds involving residues Val33O-Leu78N and Leu35N-Leu78O on strands β3 and β6 of the β-barrel (**figure 5.8**). This β-sheet extends around the whole core of the trimer forming a β-barrel of 9 strands, populated with hydrophobic residues: Leu35 (β3); Leu78, Val79, Ala80 (β6); and Leu69, Pro70, Pro72, Val73, Phe74 (C-terminal extension of β6). The β-sheet hydrogen bonds are reinforced by ion pair interactions between Asp11 and Arg34 across the monomer-monomer interface and by additional hydrogen bonds involving Thr13 and Arg32. A number of close, inter-subunit, hydrophobic contacts exist between the aromatic residues Tyr38-Pro70, Tyr24-Phe74, Phe74-Phe74, and Phe74-Pro75. All interactions were identified by the CONTACT program within the CCP4 suite, and subsequently checked manually using COOT (**table 5.4**).

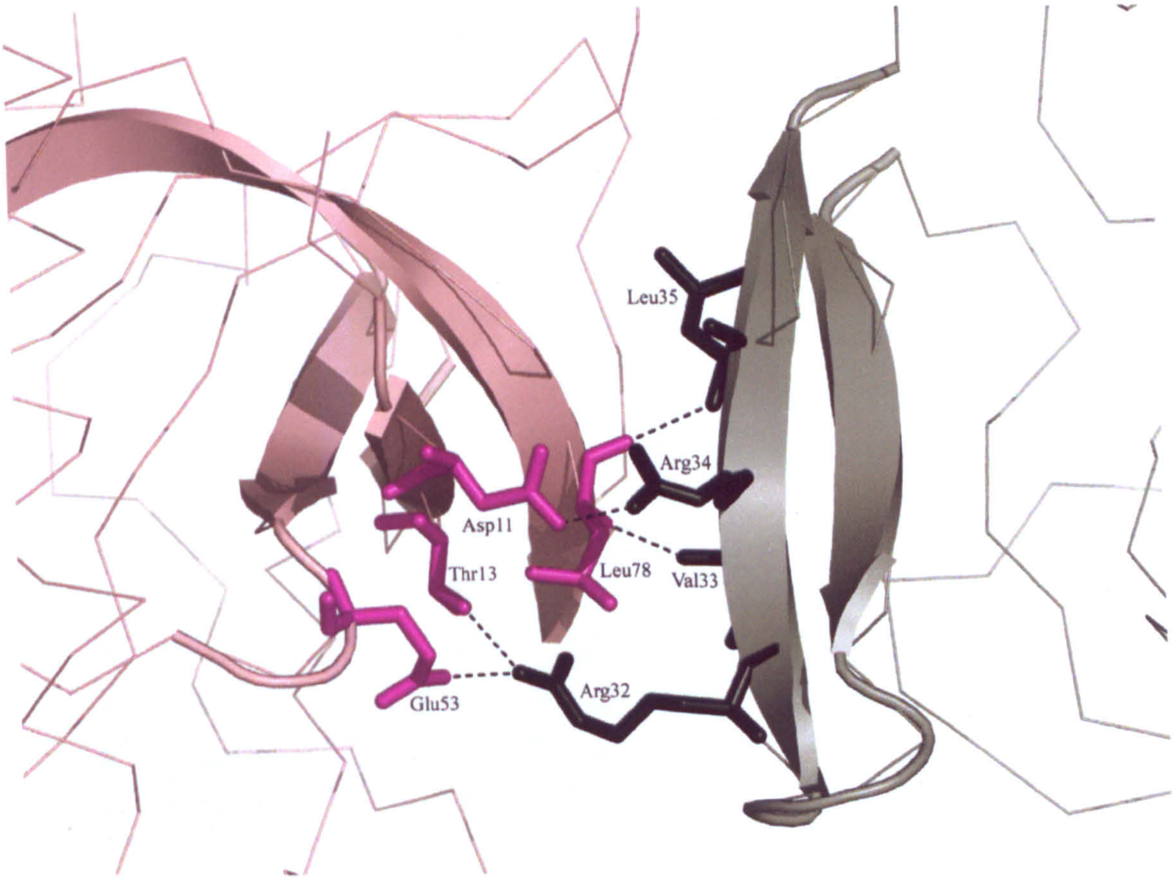


Figure 5.8: The hydrogen bonds which stabilise the Mtb-PPase trimer, shown here for chains A (pink) and C (grey). Relevant residues are shown as sticks and β-sheets as arrows.

Chain	Residue	Chain	Residue	Interaction	Distance (Å)
A	Asp11 Oδ1	B	Arg34 Nη2	Ion-pair	2.44
	Oδ2		Nη1		3.30
	Oδ2		Nη2		2.78
A	Thr13 Oγ1	B	Arg32 Nη2	H-bond	2.70
A	Glu53 Oε1	B	Arg32 Nη2	H-bond	2.50
A	Pro70 Cγ	B	Tyr38 Cδ2	Hydrophobic contact	3.43
A	Phe74 Cδ1	B	Tyr24 CZ	Hydrophobic contact	3.77
A	Phe74 Cε2	B	Phe74 Cδ2	Hydrophobic contact	3.83
A	Phe74 Cε1	B	Pro75 Cδ	Hydrophobic contact	3.88
A	Leu78 N	B	Val33 O	H-bond	2.75
A	Leu78 O	B	Leu35 N	H-bond	2.85
A	Arg32 Nη2	C	Thr13 Oγ1	H-bond	2.65
A	Val33 O	C	Leu78 N	H-bond	2.81
A	Arg34 Nη1	C	Asp11 Oδ2	Ion-pair	3.53
	Nη2		Oδ1		2.58
	Nη2		Oδ2		3.05
A	Leu35 N	C	Leu78 O	H-bond	2.89
A	Phe74 Cδ2	C	Phe74 Cε2	Hydrophobic contact	3.76
A	Pro75 Cγ	C	Phe74 Cδ1	Hydrophobic contact	3.92
B	Asp11 Oδ1	C	Arg34 Nη2	Ion-pair	2.54
	Oδ2		Nη1		3.47
	Oδ2		Nη2		2.90
B	Thr13 Oγ1	C	Arg32 Nη2	H-bond	2.70
B	Phe74 Cδ1	C	Tyr24 CZ	Hydrophobic contact	3.62
B	Phe74 Cε2	C	Phe74 Cδ2	Hydrophobic contact	3.74
B	Phe74 Cδ1	C	Pro75 Cγ	Hydrophobic contact	3.76
B	Leu78 N	C	Val33 O	H-bond	2.75
B	Leu78 O	C	Leu35 N	H-bond	2.75

Table 5.4: Mtb-PPase intra-trimer interactions, measured by CONTACT in CCP4 (Murshudov *et al.*, 1997).

The non-crystallographic trimers are packed into hexamers, related by twofold crystallographic symmetry axes (**figure 5.9**). Hydrogen bonds are formed between symmetry-related subunits A-C and B-B at residues His128Nε2 and Asp135Oδ2 (3.05 Å). Homologous interactions exist in the E-PPase hexamer (Kankare *et al.*, 1996 and Sivula *et* *al.*, 1999). The lack of strong interactions within the hexamer reflects the ease with which dissociation to trimers occurs at low pH (Schreier, 1980). The overall surface area buried in the intra-trimer interface is 552.2 Å^2.

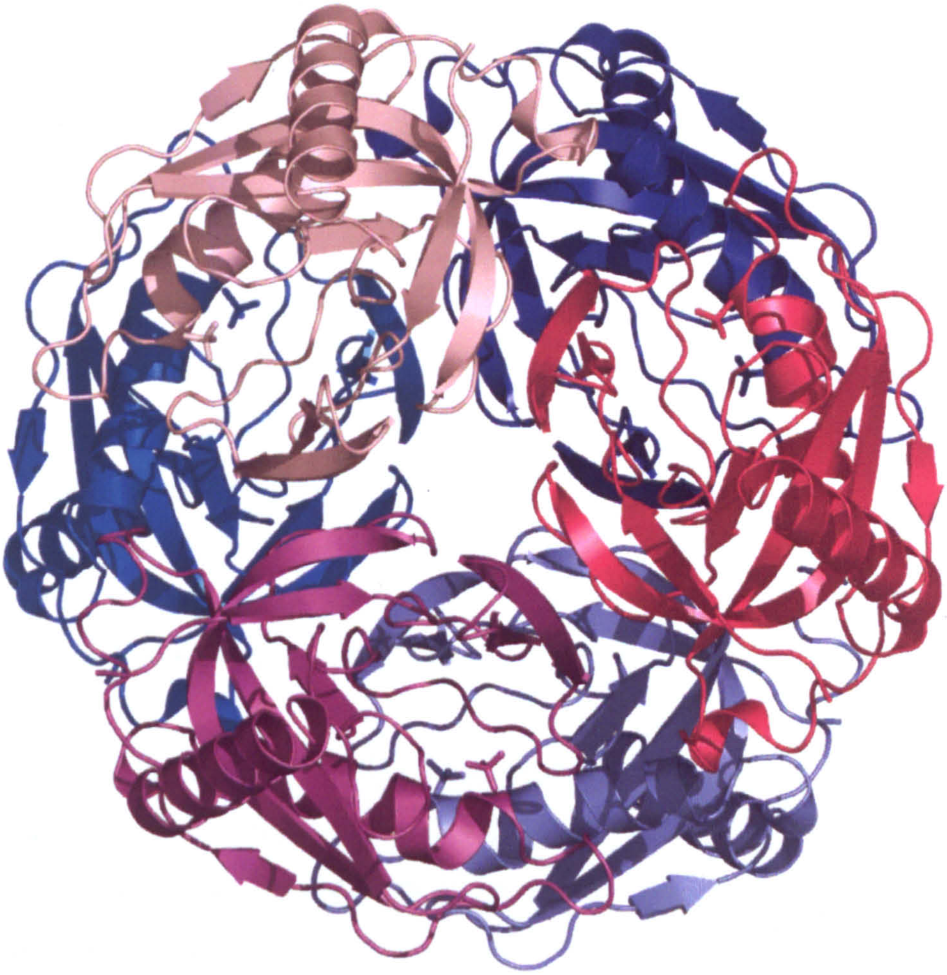


Figure 5.9: The Mtb-PPase hexamer, generated by applying symmetry operations to the crystallographic trimer. The blue/grey structure represents the symmetrical trimer: chain D (dark blue); E (light blue); and F (grey); generated from the pink/purple trimer: chain A (deep pink); B (purple); and C (light pink). Secondary structural elements are represented as cartoons and phosphate groups as sticks.

5.8.3 The active site

Conserved residues which form the type I PPase active site as described in section 1.2.7a, are also conserved in Mtb-PPase. The deep Mtb-PPase active site is formed by residues located between the C-terminal end of α1 and the exterior of the β-barrel, and contains the 13 residues involved in either substrate/metal binding or in catalysis.

The active site residues in Mtb-PPase are: Glu15 (located within strand β1); Lys23 (β2); Glu25 (β2); Arg37 (β3); Tyr49 (β4); Asp64 (β5); Asp96 (β7); and Lys98 (β7). Residues located in the loops between the β-barrel strands are: Asp59 (β4-β5); Asp61 (β4-β5); and Asp91 (β6-β7). Residues Tyr133 and Lys134 are located on the C-terminal of α1 (Tyr133) and on the loop extending from the α1 C-terminus (Lys134). The 13 active site residues in each Mtb-PPase monomer can be superimposed with rms deviations of 0.07 Å (chains A & B), 0.08 Å (chains A & C), and 0.07 Å (chains B & C).

5.8.3a The substrate/product and metal binding sites

As described in section 1.2.4, PPases require metal ions such as Mg^{2+}, Zn^{2+}, and Mn^{2+} for activity. Although Mtb-PPase was not crystallised in the presence of any such metal, the availability of magnesium ions during cell-free expression or cobalt during affinity-chromatography may have been sufficient to incorporate into the protein. To determine whether magnesium ions were present within the electron density, it was first necessary to identify active site residues. This was performed by sequence comparison with Mtb-PPase homologues of known structure, and is shown in **figure 5.12** (section 5.8.4a).

Examination of the electron density in the region of these residues did identify potential magnesium/cobalt sites, however upon further investigation these were found to not be genuine. Firstly, the distances between the "ion" and surrounding residues were predominantly longer than 3 Å. Secondly, the residues listed in **table 5.5**, do not correspond with those directly involved in metal binding in Y-PPase or E-PPase (Glu25 and Asp59, 64, 91, and 96, Mtb-PPase numbering) (Harutyunyan *et al.*, 1996 and 1997, Heikinheimo *et al.*, 1996, and Samygina *et al.*, 2001). Finally, they do not provide the negative charge required for such metal coordination. Instead, the positively charged residues surrounding this unknown region of density, suggest a phosphate-binding site (**figure 5.10**). This corresponds with data from the Y-PPase structure, where homologous residues of those listed in **table 5.5** are involved in phosphate binding (Heikinheimo *et al.*, 1996). Furthermore, residues homologous to Lys23 and Arg37 were proposed (Salminen *et al.*, 1995), and later identified (Samygina *et al.*, 2001), as product/substrate binding regions in E-PPase. Modelling of water into this region of density yielded a average B-factor of 6.61 $Å^2$, which is significantly lower than the overall B-factor of 19.1 $Å^2$ (overall solvent B-factor 32.1 $Å^2$), suggesting that water does not belong in this position either.

Although it is was not possible to distinguish between phosphate (present in the cell-free reaction solution in the form of 80 mM creatine phosphate and in the protein storage buffer as 50 mM NaH_2PO_4) and sulphate (present in the crystallisation solution in the form of 2M ammonium sulphate) in the electron density maps at this resolution, one phosphate molecule per monomer was modelled (**figures 5.10 to 5.11**), coordinated to the basic residues Arg37 and Tyr133 (**table 5.5**).

Homologous residues in Y-PPase (Harutyunyan *et al.*, 1996) and E-PPase (Samygina, 2001) have been found to bind to phosphate, denoted site P1, however no density was found at the P2 location (Lys23 and Tyr49). Whilst the possibility exists that a sulphate group may be present (from the crystallisation medium), modelling this group into the structure results in significantly elevated B-factors (59.2 $Å^2$ for sulphate and 30.6 $Å^2$ for phosphate), suggesting this may not be the case. Potentially, the phosphate may have incorporated strongly enough during cell-free synthesis and subsequent storage in phosphate buffer, to prevent substitution of sulphate during crystallisation.

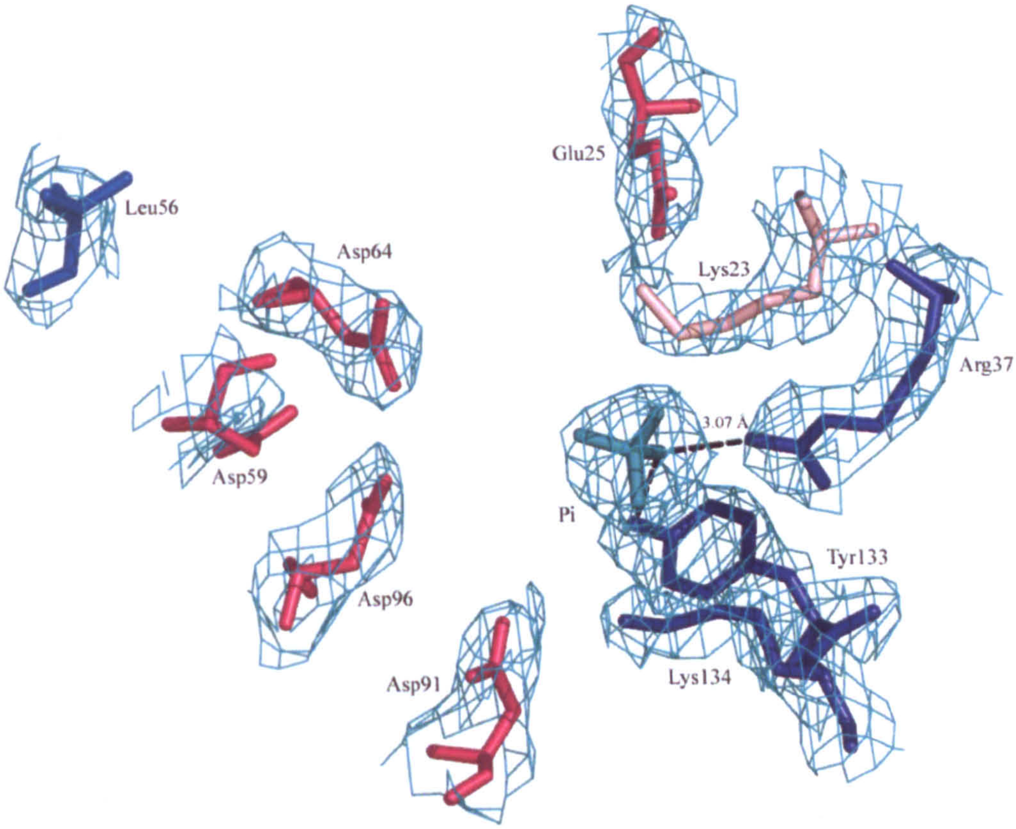


Figure 5.10: The Mtb-PPase active site (chain A) with phosphate modelled into the region of unknown density. P1 binding sites in Y-PPase are shown in dark blue, P2 sites in light blue, metal coordination regions in deep pink, and other active site residues in light pink (Heikinheimo *et al.*, 1996). The 2Fo-Fc electron density map is shown, contoured to 1.0 rms.

Chain	Residue	Phosphate ligand	H-bond bond distance (Å)
A	Arg37 Nη2	O1	3.07
	Arg37 Nη2	O4	2.62
	Tyr133 Oη	O1	2.62
B	Arg37 Nη2	01	2.96
	Arg37 Nη2	04	2.65
	Tyr133 Oη	O1	2.57
C	Arg37 Nη2	O1	3.05
	Arg37 Nη2	O4	2.68
	Tyr133 Oη	O1	2.64

Table 5.5: Distances between the modelled product (phosphate) binding sites and their Mtb-PPase ligands for each unit of the trimer. Distances measured manually using COOT (Emsley and Cowtan, 2004).

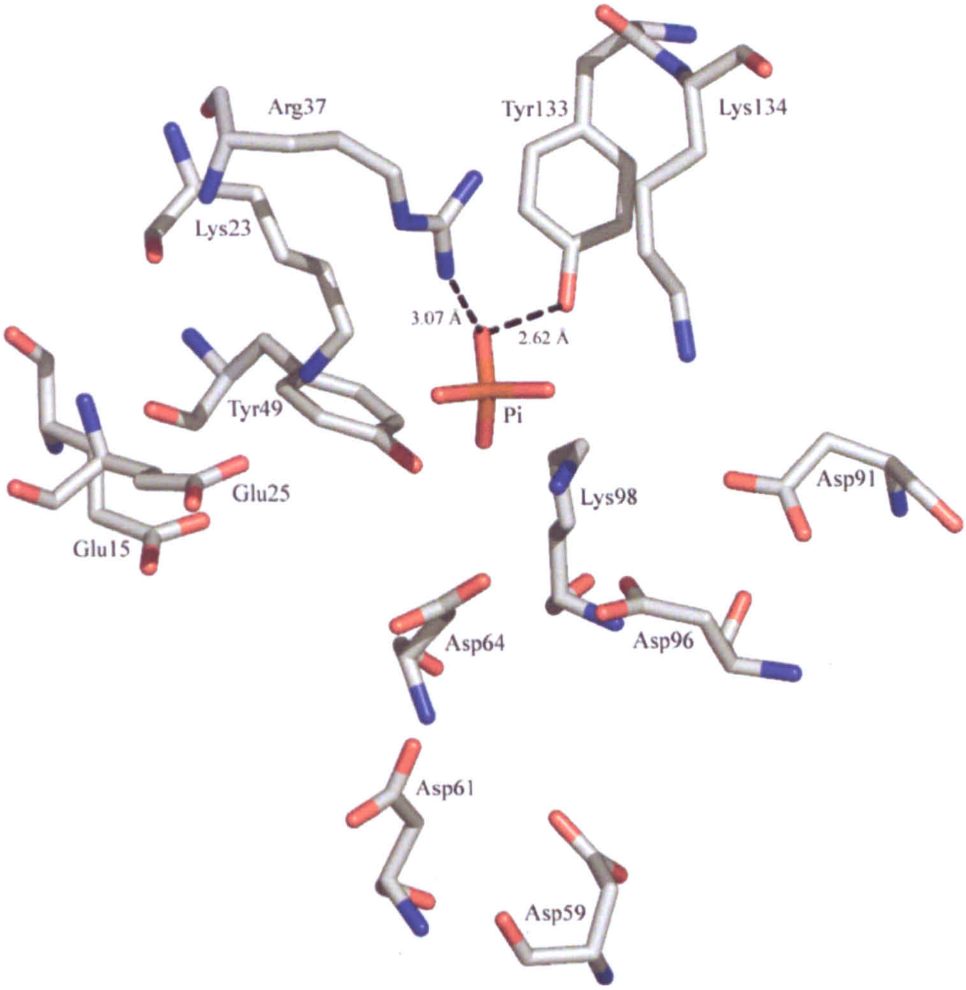


Figure 5.11: The active site region of Mtb-PPase. Carbon is represented in grey; oxygen in red; nitrogen in blue; and phosphorous in orange. Distances between the phosphate group oxygens and its ligands are labelled (chain A).

5.8.4 Comparisons with type I inorganic pyrophosphatases

5.8.4a Primary structure

The Rv3628 protein sequence was obtained from the TBSGC and used to search for homologues within both the UniProt and the PDB databases, using the NCBI BLAST2 blastp software (http://www.ebi.ac.uk/blastall/index.html). All of the twenty most similar sequences, identified from the UniProt database, were from bacterial sources (**table 5.7**). The PDB search identified six unique PPase structures, one of which, yeast PPase, shares just 25% sequence identity (**table 5.6**). This illustrates the low sequence identity between prokaryotic and eukaryotic PPases. Despite this, alignment of these six sequences identified 17 conserved residues (**figure 5.12**), including the 13 active site residues required for catalytic activity in type I PPases (see section 1.2.7a).

Although an X-ray crystal structure is not available for human PPase (H-PPase, NCBI gi accession code: 33150672), a sequence comparison was carried out (**figure 5.13**). Despite sharing only 25 % identity with Mtb-PPase, the active site residues were strictly conserved. A crystal structure of human PPase is therefore a priority in order to identify structural differences, which could lead to the use of Mtb-PPase as a target for rational drug design.

Species	**PDB ID**	**Score**	**Length**	**Sequence identity (%)**
Pyrococcus horikoshii	1UDE	392	195	49
Thermus thermophilus	2PRD	391	174	51
Sulfolobus acidocaldarius	1QEZ	390	173	45
Pyrococcus furiosus	1TWL	377	186	47
Escherichia coli	1FAJ	358	175	45
Saccharomyces cerevisiae	1WGJ	112	286	25

Table 5.6: Output from an NCBI BLAST blastp PDB database search of the Rv3628 gene product, Mtb-PPase. Data taken from an output of the fifty highest scoring sequences (all inorganic pyrophosphatases). Sequences listed more than once were removed from the output.

Species	**Protein**	**Length**	**Sequence identify (%)**
Mycobacterium tuberculosis	Inorganic pyrophosphatase	162	100
Mycobacterium bovis	Inorganic pyrophosphatase	162	100
Mycobacterium leprae	Inorganic pyrophosphatase	162	89
Mycobacterium paratuberculosis	Inorganic pyrophosphatase	162	86
Rhodococcus sp.	Inorganic diphosphatase	163	82
Mycobacterium flavescens	Inorganic pyrophosphatase	161	80
Mycobacterium vanbaalenii (PYR-1)	Inorganic diphosphatase	161	79
Mycobacterium sp. (JLS)	Inorganic diphosphatase	162	75
Mycobacterium sp. (MCS)	Inorganic diphosphatase	162	75
Nocardia farcinica	Putative inorganic pyrophosphatase	163	76
Corynebacterium diphtheriae	Inorganic pyrophosphatase	158	68
Streptomyces coelicolor	Inorganic pyrophosphatase	163	68
Acidothermus cellulolyticus (11B)	Inorganic diphosphatase	161	66
Kineococcus radiotolerans (SRS30216)	Inorganic diphosphatase	173	67
Streptomyces avermitilis	Putative inorganic pyrophosphatase	163	65
Thermobifida fusca	Inorganic diphosphatase	171	67
Nocardioides sp. (JS614)	Inorganic diphosphatase	163	64
Corynebacterium efficiens	Inorganic pyrophosphatase	158	65
Corynebacterium jeikeium (K411)	Inorganic pyrophosphatase	160	64
Corynebacterium glutamicum (Brevibacterium flavum)	Inorganic pyrophosphatase	158	62

Table 5.7: Output from an NCBI BLAST blastp UniProt database search of the Rv3628 gene product, Mtb-PPase. Data taken from an output of the twenty highest scoring sequences. Sequences listed more than once were removed from the output.

```
                                                                     β1          β2
                                                                   ________     _____
Mtb-PPase                 --------------------------------GS-SGSSGMQFDVTIEIPKGQR-NKYEV 26
T-PPase (2PRD 51 %)       --------------------------ANLKSLPV-GDKAPEVVHMVIEVPRGSG-NKYEY 32
Pho-PPase (1UDE_A 49 %)   --------HHHHHHSSGLVPRGSHMMNPFHDLEP-GPNVPEVVYALIEIPKGSR-NKYEL 50
Pfu-PPase (1TWL_A 47 %)   -------AHHHHHHGS----------NPFHDLEP-GPDVPEVVYAIIEIPKGSR-NKYEL 41
E-PPase (1FAJ 45 %)       ---------------------------SLLNVPA-GKDLPEDIYVVIEIPANADPIKYEI 32
S-PPase (1QEZ_A 45 %)     -----------------------------MKLSP-GKNAPDVVNVLVEIPQGSN-IKYEY 29
Y-PPase (1WGJ_A 25 %)     TYTTRQIGAKNTLEYKVYIEKDGKPVSAFHDIPLYADKENNIFNMVVEIPRWTN-AKLEI 59
                                       β3                  β4
                                 ____________            _____
Mtb-PPase                 DHET--GRVRLD------RYLYTPM---AYPTDYGFIEDTLGD------------DGDPL 63
T-PPase                   DPDL--GAIKLD------RVLPGAQ---FYPGDYGFIPSTLAE------------DGDPL 69
Pho-PPase                 DKET--GLLKLD------RVLYTPF---HYPVDYGIIPRTWYE------------DGDPF 87
Pfu-PPase                 DKKT--GLLKLD------RVLYSPF---FYPVDYGIIPRTWYE------------DDDPF 78
E-PPase                   DKES--GALFVD------RFMSTAM---FYPCNYGYINHTLSL------------DGDPV 69
S-PPase                   DDEE--GVIKVD------RVLYTSM---NYPFNYGFIPGTLEE------------DGDPL 66
Y-PPase                   TKEETLNPIIQDTKKGKLRFVRNCFPHHGYIHNYGAFPQTWEDPNVSHPETKAVGDNDPI 119
                            β5                β6            β7           α3     α4  α1
                          ____         ______________  __________     ______   _____--
Mtb-PPase                 DALVLLPQPVFPGVLVAARPVGMFRMVDEHGGDDKVLCVPAG---DPRWDHVQDIGDVPA 120
T-PPase                   DGLVLSTYPLLPGVVVEVRVVGLLLMEDEKGGDAKVIGVVAE---DQRLDHIQDIGDVPE 126
Pho-PPase                 DIMVIMREPTYPLTIIEARPIGLFKMIDSGDKDYKVLAVPVE---DPYFKDWKDISDVPK 144
Pfu-PPase                 DIMVIMREPVYPLTIIEARPIGLFKMIDSGDKDYKVLAVPVE---DPYFKDWKDIDDVPK 135
E-PPase                   DVLVPTPYPLQPGSVIRCRPVGVLKMTDEAGEDAKLVAVPHSK-LSKEYDHIKDVNDLPE 128
S-PPase                   DVLVITNYQLYPGSVIEVRPIGILYMKDEEGEDAKIVAVPKDK-TDPSFSNIKDINDLPQ 125
Y-PPase                   DVLEIGETIAYTGQVKQVKALGIMALLDEGETDWKVIAIDINDPLAPKLNDIEDVEKYFP 179
                                                        β8          α2
                          -------------            ____________---------------
Mtb-PPase                 FELDAIKHFFVHYKDLEP--GKFVKAAD----WVDRAEAEAEVQRSVERFKAGTH----- 169
T-PPase                   GVKQEIQHFFETYKALEAKKGKWVKVTG----WRDRKAALEEVRACIARYKG-------- 174
Pho-PPase                 AFLDEIAHFFKRYKELEG---KEIIVEG----WEGAEAAKREILRAIEMYKEKFGKKE-- 195
Pfu-PPase                 AFLDEIAHFFKRYKELQG---KEIIVEG----WEGAEAAKREILRAIEMYKEKFGKKE-- 186
E-PPase                   LLKAQIAHFFEHYKDLEK--GKWVKVEG----WENAEAAKAEIVASFERAKNK------- 175
S-PPase                   ATKNKIVHFFEHYKELEP--GKYVKISG----WGSATEAKNRIQLAIKRVSGGQ------ 173
Y-PPase                   GLLRATNEWFRIYKIPD---GKPENQFAFSGEAKNKKYALDIIKETHDSWKQLIAGKSSD 236

Y-PPase                   SKGIDLTNVTLPDTPTYSKAASDAIPPASLKADAPIDKSIDKWFFISGSV           286
```

Figure 5.12: Amino acid sequence alignment (Thompson *et al.*, 2000) of Mtb-PPase and type I PPases from *Thermus thermophilus* (T-PPase); *Pyrococcus horikoshii* (Pho-PPase); *Pyrococcus furiosus* (Pfu-PPase); *Escherichia coli* (E-PPase); *Sulfolobus acidocaldarius* (S-PPase); and *Saccharomyces cerevisiae* (Y-PPase). Sequence identity and PDB ID are shown in parenthesis. Conserved residues in all known soluble type I PPases are shown in boldface (Sivula *et al.*, 1999). Catalytically essential and phosphate/metal-binding residues are shown in red (Y-PPase, Heikinheimo, 1996). Secondary structural information for Mtb-PPase is shown above the sequence, as determined by PROMOTIF (Hutchinson and Thornton, 1996). Novel histidine active site residues in Mtb-PPase are highlighted in yellow (Tammenkoski *et al.*, 2005).

```
Mtb-PPase         ----------------------GSSG------------SSGMQFDVTIEIPKGQRNKYEV 26
H-PPase (25 %)    MSGFSTEERAAPFSLEYRVFLKNEKGQYISPFHDIPIYADKDVFHMVVEVPRWSNAKMEI 60

Mtb-PPase         DHET--GRVRLD------RYLYTPM---AYPTDYGFIEDTLGD------------DGDPL 63
H-PPase           ATKDPLNPIKQDVKKGKLRYVANLFPYKGYIWNYGAIPQTWEDPGHNDKHTGCCGDNDPI 120

Mtb-PPase         DALVLLPQPVFPGVLVAARPVGMFRMVDEHGGDDKVLCVPAG---DPRWDHVQDIGDVPA 122
H-PPase           DVCEIGSKVCARGEIIGVKVLGILAMIDEGETDWKVIAINVDDPDAANYNDINDVKRLKP 180

Mtb-PPase         FELDAIKHFFVHYKDLEP--GKFVKAADWVDRAEAEAEVQRSVERFKAGTH--------  169
H-PPase           GYLEATVDWFRRYKVPD---GKPENEFAFNAEFKDKDFAIDIIKSTHDHWKALVTKKTN  236

H-PPase           GKGISCMNTTLSESPFKCDPDAARAIVDALPPPCESACTVPTDVDKWFHHQKN        289
```

Figure 5.13: Amino acid sequence alignment (Thompson *et al.*, 2000) of Mtb-PPase and PPase from *Homo sapiens* (H-PPase, 25 % sequence identity). Conserved residues in all known soluble type I PPases are shown in boldface (Sivula *et al.*, 1999). Catalytically essential and phosphate/metal-binding residues are shown in red (Y-PPase, Heikinheimo, 1996).

5.8.4b Overall fold

No significant variation in the overall monomeric fold of Mtb-PPase was identified in comparison with existing prokaryotic PPase structures of comparable size (178 ± 17 residues). When superimposed, the core tertiary structures of all type I prokaryotic PPases are virtually indistinguishable (**figure 5.14**). Cα atoms of the Mtb-PPase structure (chain A) were superimposed onto the structures of several PPases giving rms deviations of: 1.04 Å, T-PPase (Teplyakov *et al.*, 1994); 0.94 Å, Pho-PPase (Liu *et al.*, 2004); 0.93 Å, E-PPase CaPPi (Samygina *et al.*, 2001); 0.87 Å, S-PPase Mg (Leppanen *et al.*, 1999); 0.85 Å, E-PPase apoenzyme (Kankare *et al.*, 1996); and 0.80 Å, Pfu-PPase (Zhou *et al.*, 2006).

Aligning the Cα atoms (chain A) of the thirteen active site residues alone gave rms deviations of: 2.07 Å, E-PPase (apoenzyme, 2.15 Å); 1.94 Å, Pfu-PPase (apoenzyme, 2.2 Å); 1.60 Å, T-PPase (SO_4, 2.0 Å); 0.89 Å, Pho-PPase (apoenzyme, 2.7 Å); and 0.82 Å, E-PPase (CaPPi, 1.2 Å). Variation in the orientation of the apoenzymes (Pfu-, Pho-, and E-PPase) with Mtb-PPase is a partial result of the Mtb-PPase active site adopting a different conformation in the PO_4-bound state. Conversely, alternate conformations of metal-binding residues occur in the calcium-inhibited structure of E-PPase. Aligning the active site residues of Mtb-PPase and T-PPase gave a larger rms deviation value than might be expected given that both structures have PO_4/SO_4 bound.

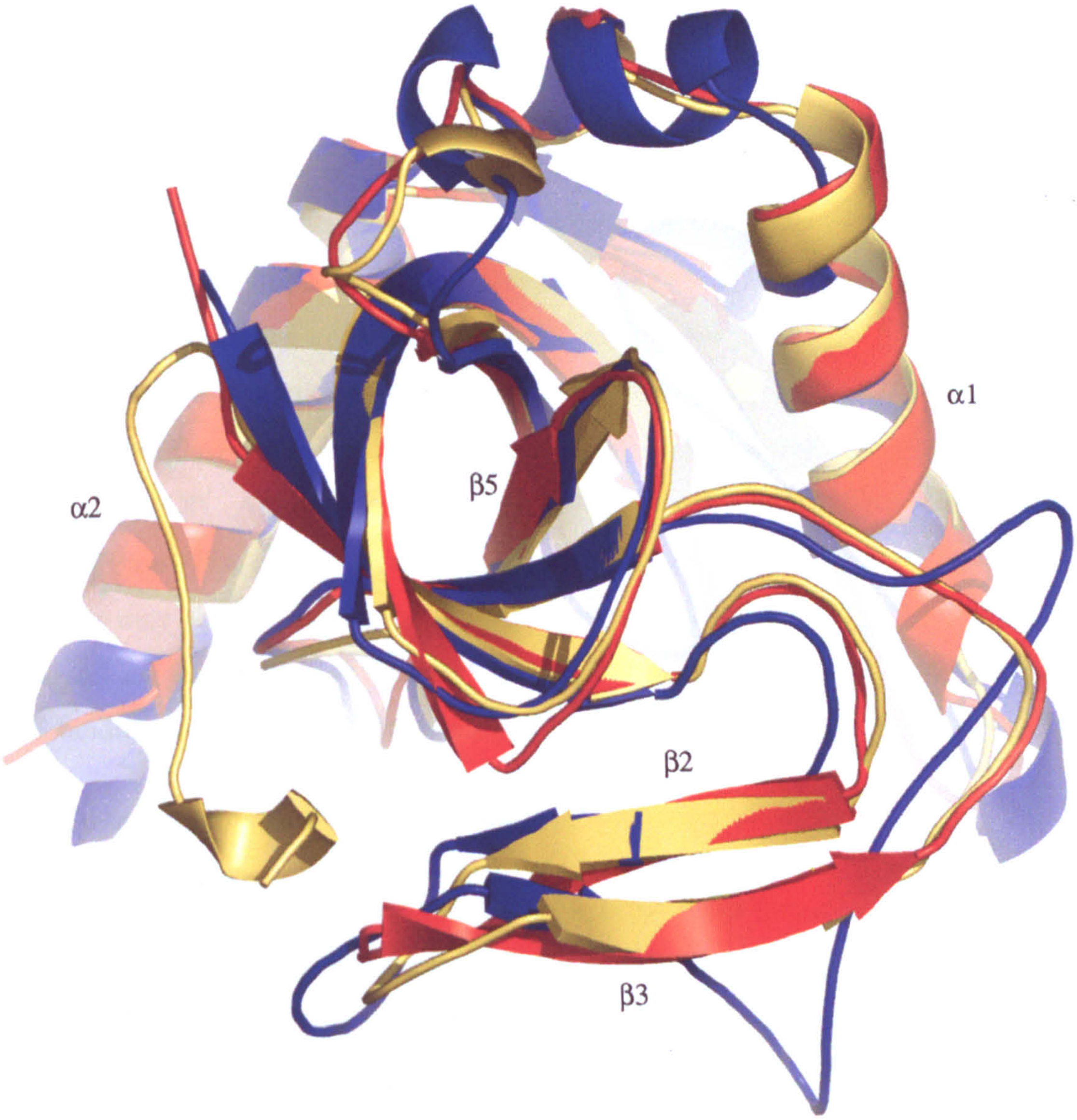


Figure 5.14: Superimposition of three PPase structures (chains A): Mtb-PPase PO_4 (represented in red); E-PPase apoenzyme (yellow) (Kankare *et al.*, 1996); and the Y-PPase $(MnPi)_2$ core, residues 41-230 (blue) (Heikinheimo *et al.*, 1996), highlights the striking similarity in the overall fold of type I PPases, despite Y-PPase sharing only 25 % sequence identity with Mtb-PPase. The highly distorted, five stranded β-barrel and the two large α-helices are clearly conserved within these structures.

5.8.4c Oligomeric form

The oligomeric form for prokaryotic type I PPases, described in section 1.2.7c, remains conserved for Mtb-PPase, despite intra-trimer interactions involving poorly-conserved residues. The Cα atoms of the Mtb-PPase non-crystallographic trimer were superimposed onto the S-PPase and Pho-PPase trimers, with rms deviations of 1.30 and 1.49 Å, respectively. The typical hexameric arrangement, formed from two dimers, is also conserved in Mtb-PPase (**figure 5.9**).

5.8.4d Active site

The active site cavity for PPases which have metal and/or phosphate bound tend to adopt a tighter conformation. In particular, residues known to form interactions with these ligands are orientated differently so as to allow for closer contacts (Samygina *et al.*, 2001). Comparison of these orientations, by aligning individual active site residues (Cα atoms) from different PPases, was used to further substantiate the role of these residues in ligand binding (**table 5.8**). Although active site residues are conserved between E-PPase (CaPPi) and Mtb-PPase (Pi), conformation of some of these residues differ quite significantly. Whilst this may be attributed to the different resolution at which X-ray data were collected (1.2 Å and 2.7 Å, E-PPase and Mtb-PPase respectively), it seems sensible to assume that ligand binding would induce a structural shift within the active site.

5.8.4e Comparison with T-PPase, in complex with sulphate

Conformational changes within the active sites of Mtb-PPase and other homologous structures may be as a result of substrate/product binding, metal incorporation, or due to differences in resolution. As described in section 5.8.4b, aligning the Cα atoms (chain A) of Mtb-PPase and T-PPase (PDB ID: 2PRD) active site residues showed a greater variation than expected, considering the presence of one phosphate/sulphate group in each monomer. Firstly, the orientation of the T-PPase SO_4 differs from the Mtb-PPase PO_4 group, with the Mtb-PPase phosphate being slightly more centralised within the active site, resulting in a greater distance between itself and Lys134 Nζ (0.40 Å) (**figure 5.15**). However, this is likely to be partly due to the different resolution at which the two data sets were collected.

Of the three P1 binding residues, the Arg37 and Lys134 side chains vary the most between the two structures. Mtb-PPase Arg37 adopts a flipped conformation at atom Nε, which results in Nη1 facing away from the modelled phosphate. A similar shift is also seen for Lys134, whereby the Mtb-PPase Cδ and Cε atoms flip, resulting in a greater distance between Nζ and the phosphate. Finally, the main chain of Asp96 adopts a more extended conformation in T-PPase, with the Oδ1 atom orientated at approximately 90 ° to the corresponding atom in Mtb-PPase. Again, these differences may be due to the improved resolution at which the T-PPase data were collected.

Table 5.8: Root mean square deviations (Å) from superimposing the 13 conserved active site residues of Mtb-PPase with equivalent residues in homologous PPases, using PYMOL (DeLano Scientific). Ligands bound to each PPase and the resolution at which data were collected are shown in parentheses. PPi and metal-binding residues are highlighted in red and boldface, respectively (Heikinheimo *et al.*, 1996, Harutyunyan *et al.*, 1996, and Samygina *et al.*, 2001). Where multiple subunits exist, chain A was used for the alignment.

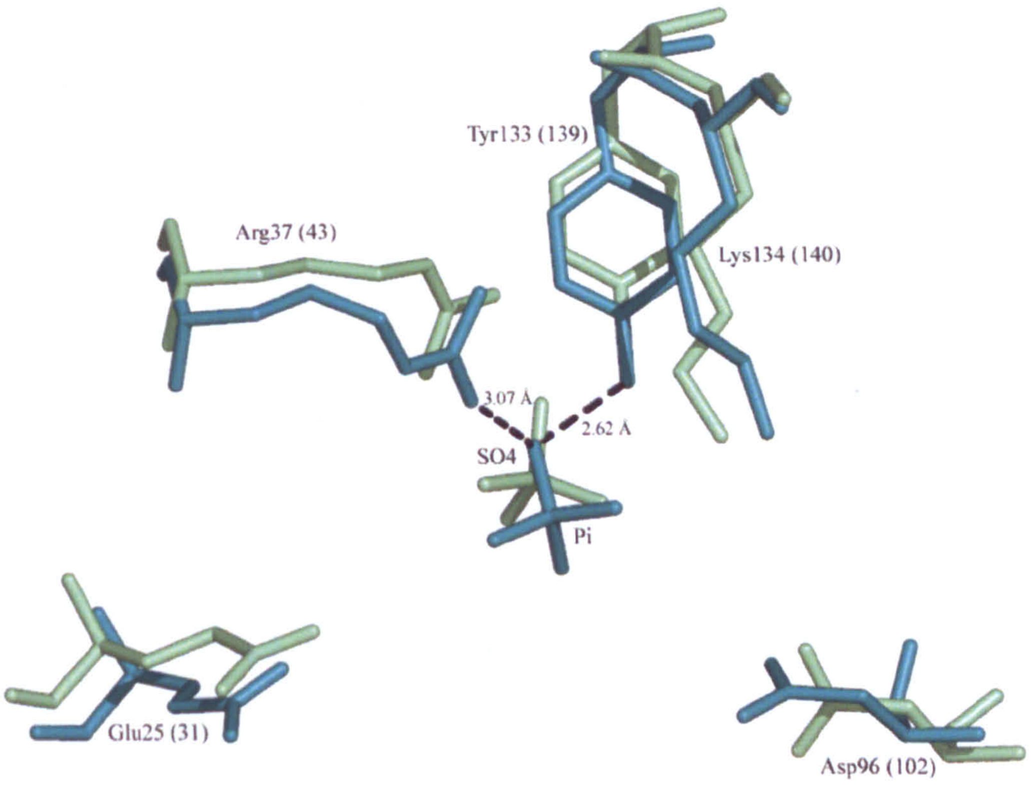


Figure 5.15: Superimposition of Mtb-PPase chain A (shown in teal) and T-PPase SO_4 (Teplyakov *et al.*, 1994) (shown in light green) active site residues with high rms deviations between the two structures. Interactions with Mtb-PPase residues and the modelled phosphate are shown as black broken lines.

5.8.4f Comparison with E-PPase, in complex with its natural inhibitor calcium

Binding of the natural inhibitor, Ca^{2+}, to E-PPase at the metal binding sites, M1-3, induced a structural shift similar to that of Mn^{2+} binding (Samygina *et al.*, 2001). Aligning the Cα atoms (chain A) of Mtb-PPase and E-PPase (CaPPi) (PDB ID: 1I6T) active site residues identified seven with high rms deviations (> 0.38 Å) (**table 5.8**). The most prominent of these deviations are in residues Glu25 and Asp96 (metal binding residues), which align with corresponding residues in E-PPase (CaPPi) with rms deviations of > 0.8 Å (**table 5.8**). The Glu25 in Mtb-PPase is positioned 1.57 Å (Cδ to P atoms) further away from P1 than in the E-PPase structure (**figure 5.16**). The side chain is flipped from Cγ onwards, with a distance of 2.42 Å between the two positions of the Cγ atom. The E-PPase conformation of these residues is also seen in the Y-PPase $(MnPi)_2$ structure, emphasising the role of these residues in metal binding. In the E-PPase structure, Asp96 is orientated such that both Oδ atoms have close contact with a modelled calcium, whilst in Mtb-PPase, the Oδ1 faces away from this atom.

Extension of the E-PPase Asp59 main chain towards the active site allows for a 2.75 Å closer contact between Oδ1 and the modelled calcium, than is evident within the Mtb-PPase structure. Also, an extended conformation exists for E-PPase Lys23 than in Mtb-PPase, allowing for a 1.05 Å closer contact with the P2 site, which is not present in the Mtb-PPase structure. Finally, both Nη atoms of Arg37 form contacts with the P1 site within the E-PPase structure. In Mtb-PPase, only Nη2 interacts with the phosphate group.

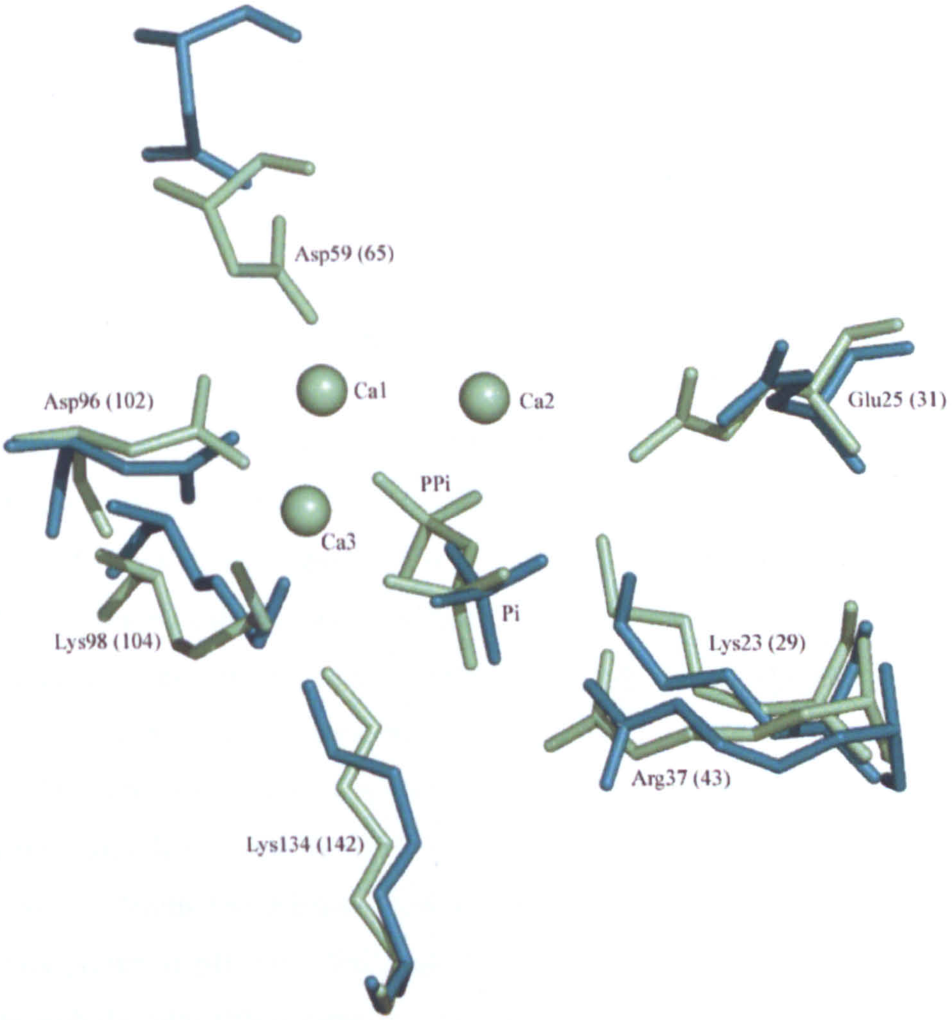


Figure 5.16: Superimposition of Mtb-PPase chain A (shown in teal) and E-PPase CaPPi (Samygina *et al.*, 2001) (shown in light green) active site residues with high rms deviations. E-PPase calcium atoms are shown as spheres.

5.8.4g Comparison with Y-PPase

Superimposing the Cα atoms of Mtb-PPase (chain A) with the Y-PPase (PDB ID: 1WGJ) core (residues 41 - 230) gave an rms deviation of 2.4 Å, whilst superimposing the 13 active site residues alone gave an rms deviation of 0.92 Å (**table 5.8**). Similar values are obtained when superimposing other prokaryotic PPases with Y-PPase (data not shown), which demonstrates the divergence of Y-PPase from prokaryotic orthologues. Y-PPase forms an elongated 286 residue subunit with N- and C- terminal extensions which form an additional β-sheet and a long β- loop, elements found to be involved in oligomeric interactions (Heikinheimo *et al.*, 1996). Aside from these differences, the central core of the Y-PPase monomer retains the tertiary fold characteristic of type I PPases, as described previously. Unlike all prokaryotic PPases, Y-PPase forms a homodimer stabilised mainly by stacking of aromatic rings. Mtb-PPase and Y-PPase share 25 % sequence similarity, with the 17 explicitly conserved residues accounting for 9 %, explaining why the rms deviation for superimposing active site residues alone is significantly lower than that for the monomer as a whole.

5.8.4h Comparison with Mtb-PPase, in space group $P6_322$

Soon after completion of the Mtb-PPase crystal structure in space group $P3_221$, another Mtb-PPase structure was deposited in the PDB, in space group $P6_322$, to a resolution of 1.3 Å (PDB ID: 1SXV, Tammenkoski *et al.*, 2005). This was crystallised in the presence of 1.7 M ammonium sulphate and 100 mM sodium acetate, pH 5.0. The structure was solved by molecular replacement using T-PPase as a starting model, with one monomer in the asymmetric unit. The structure was refined to a final R-factor of 15.4 % (R-free 16.9 %), with 238 water molecules and one sulphate group modelled. Later, another structure became available to a resolution of 1.54 Å, containing both phosphate and potassium (PDB ID: 1WCF, Benini and Wilson, to be published). This second structure was produced from crystals grown at pH 7.0. The availability of these structures allows a comparison to be made with the Mtb-PPase structure described in this chapter.

The superimposition of all chain A Cα atoms of Mtb-PPase P3$_2$21 (in this section, referred to as P3$_2$21) with Mtb-PPase P6$_3$22 pH 5.0 SO$_4$ (1SXV) and Mtb-PPase P6$_3$22 pH 7.0 K$_2$Pi (1WCF), gave rms deviations of 0.35 Å and 0.34 Å, respectively (**figure 5.17**). Superimposing the 13 active site residues alone gave rms deviations of 0.48 Å (1SXV) and 0.36 Å (1WCF). It seems likely that these differences are at least partially attributed to the improved resolution at which the two P6$_3$22 structures were collected. To determine the contribution to these differences, individual active site residues were superimposed using PYMOL (DeLano Scientific) (**table 5.9**).

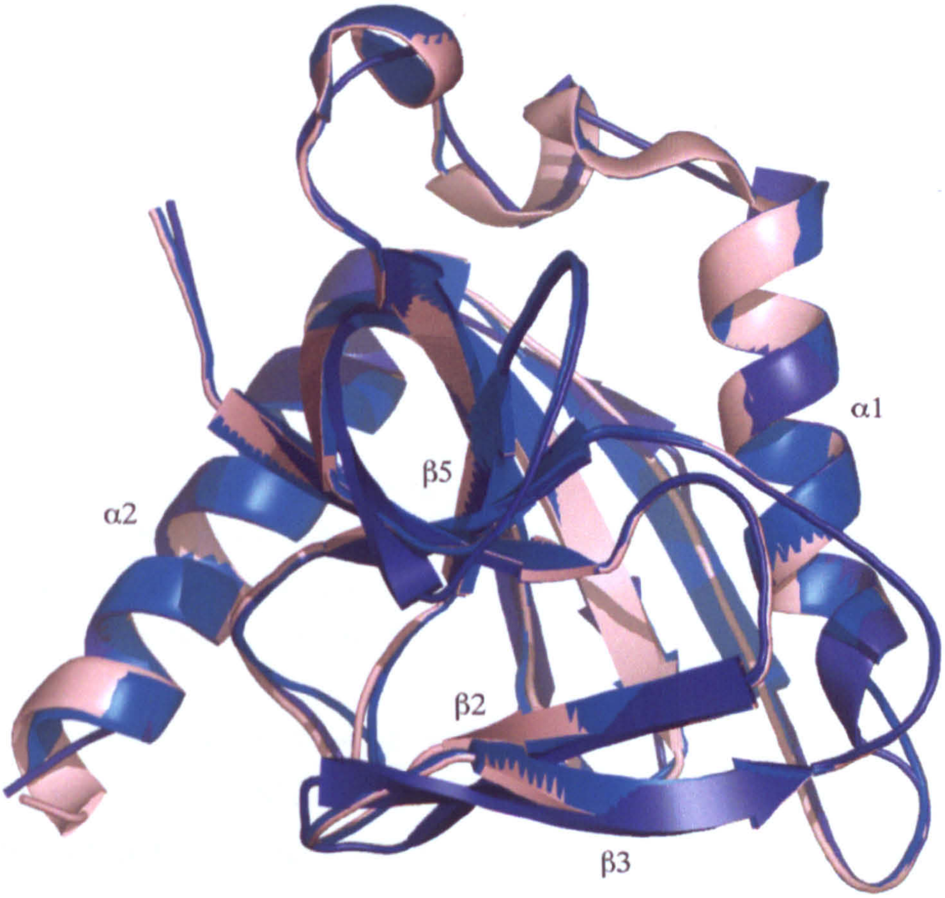


Figure 5.17: Superimposition of three Mtb-PPase structures (chains A): Mtb-PPase P3$_2$21 (represented in dark blue); Mtb-PPase 1SXV (Tammenkoski *et al.*, 2005) (light pink); and Mtb-PPase 1WCF (Benini and Wilson, to be published) (light blue).

Residue	**Rms deviation**	
	1SXV (SO_4) 1.3 Å **$P6_322$ pH 5.0**	**1WCF (K_2Pi) 1.54 Å** **$P6_322$ pH 7.0**
Glu15	0.083	0.077
Lys23	0.035	0.041
Glu25	0.004	0.048
Arg37	0.458	0.464
Tyr49	0.093	0.073
Asp59	0.511	0.607
Asp61	0.220	0.135
Asp64	0.084	0.134
Asp91	0.165	0.083
Asp96	0.662	0.800
Lys98	0.041	0.045
Tyr133	1.371	1.372
Lys134	0.220	0.154

Table 5.9: Root mean square deviations from superimposing the 13 conserved active site residues (chain A) of Mtb-PPase $P3_221$ (PO_4, pH 7.5) with Mtb-PPase 1SXV (Tammenkoski *et al.*, 2005) and Mtb-PPase 1WCF (Benini and Wilson, to be published). Ligands bound to each PPase are shown in parentheses. PPi and metal-binding residues are highlighted in red and boldface, respectively.

Of interest are the observations of Tammenkoski *et al.* (2005) of two histidine residues within the active site of 1SXV, His28 and His93, the latter of which interacts with P1. These are not found in any type I PPases, but are generally conserved within type II PPases (Fabrichniy *et al.*, 2006). His93 was not regarded with interest in the Mtb-PPase $P3_221$ structure as its adopts a less prominent position within the active site, due to an alternate conformation, which results in a significantly greater distance between itself and the modelled phosphate (**figure 5.18**). In 1SXV, there is localised extension of the main chain towards the active site, with the His93 side chain orientated such that the Nε2 atom is positioned 2.66 Å from the sulphate O4. The His93 ring within the Mtb-PPase $P3_221$ structure is angled such that the Nδ1 atom is the closest to the phosphate group. The resulting distance between the histidine Nδ1 and the phosphate O3 group is 5.41 Å, clearly too long for hydrogen bond formation. This was assumed to be a problem with the lower resolution at which the latter data were collected, however upon inspection of the 1WCF structure (Benini and Wilson, to be published), a similar conformation was observed, with a 6.12 Å distance between the histidine Nδ1 and the phosphate O4. This is unlikely to be caused by the different ions (sulphur/phosphorous) found in the structure, due to their near-identical characteristics, or by an error with the 1SXV model, due to the resolution at which data were collected. Instead this suggests a possible pH-dependence.

The two crystal structures at more neutral pH (1WCF, pH7.0 and $P3_221$, pH 7.5) did not suggest H-bond bond formation with the phosphate groups, whilst the more acidic crystal structure, 1SXV, identified a 2.66 Å interaction with a sulphate oxygen, due to protonation of the His93 (Benini and Wilson, 2004).

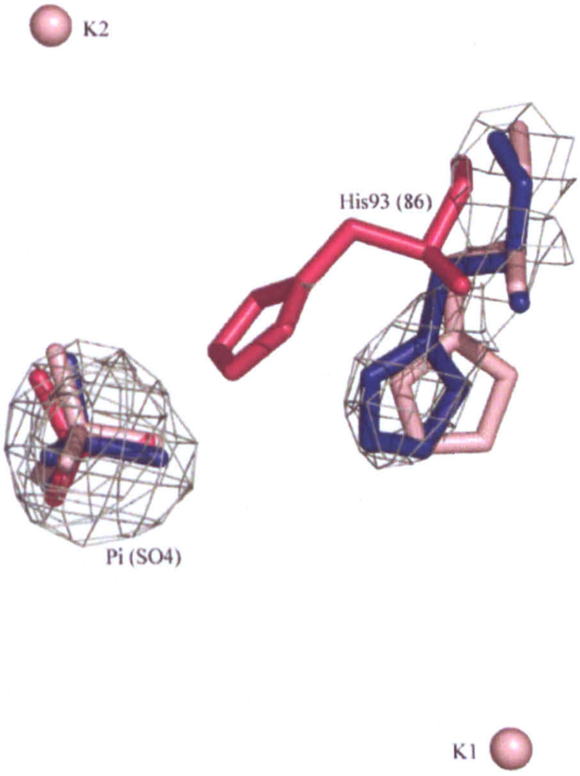


Figure 5.18: Superimposition of three Mtb-PPase His93 (86) residues (chains A) in relation to the modelled phosphate/sulphate groups. Mtb-PPase $P3_221$ (represented in dark blue); Mtb-PPase 1SXV (Tammenkoski *et al.*, 2005) (dark pink); and Mtb-PPase 1WCF (Benini and Wilson, to be published) (light pink). Mtb-PPase 1WCF potassium atoms are shown as spheres. The 2Fo-Fc electron density map for Mtb-PPase $P3_221$ is shown, contoured to 1.0 rms.

The orientation of His28 remains fairly similar between the three structures. Mutations of this residue resulted in a marked decrease in catalytic activity, with the H28K mutation resulting in a four-fold loss in the presence of magnesium (Tammenkoski *et al.*, 2005). Interestingly however, activity was increased three-fold in the presence of zinc ions, highlighting the non-essential nature of these histidines (Tammenkoski *et al.*, 2005).

The 1WCF structure contains metal ions, however examination of the structure found these two potassium atoms do not locate the activating metal binding sites typical of type I PPases. Residues involved in metal binding within type II PPases do not account for the positioning of these ions either, despite the identification of type II-like active site histidines within Mtb-PPase.

Further conformational differences of interest, within the active site residues of the three Mtb-PPase structures, are described here (**figure 5.19**). The Nε atoms of Arg37 in both Mtb-PPase $P6_322$ structures are rotated by approximately 180 ° in comparison with the $P3_221$ counterpart, resulting in both Nη atoms facing towards the phosphate/sulphate groups. In the $P3_221$ structure, only the Nη2 group makes contact with the phosphate (2.62 Å). In Y-PPase, both Nη atoms make direct contact with phosphate oxygens, suggesting this should also be the case in the $P3_221$ structure.

The main chain between Gly57 and Asp61 and around Asp96 of Mtb-PPase $P3_221$ is slightly twisted, in comparison with the Mtb-PPase $P6_322$ structures. In the 1WCF structure, the Oδ1 group of Asp96 is also 1.36 Å closer to the modelled potassium, than in the metal-free Mtb-PPase $P3_221$ structure. This residue is known to participate in metal binding, and although the position of this potassium ion does not represent a catalytic binding site, it is likely that the negative charge of the Asp96 side chain enables coordination with this atom.

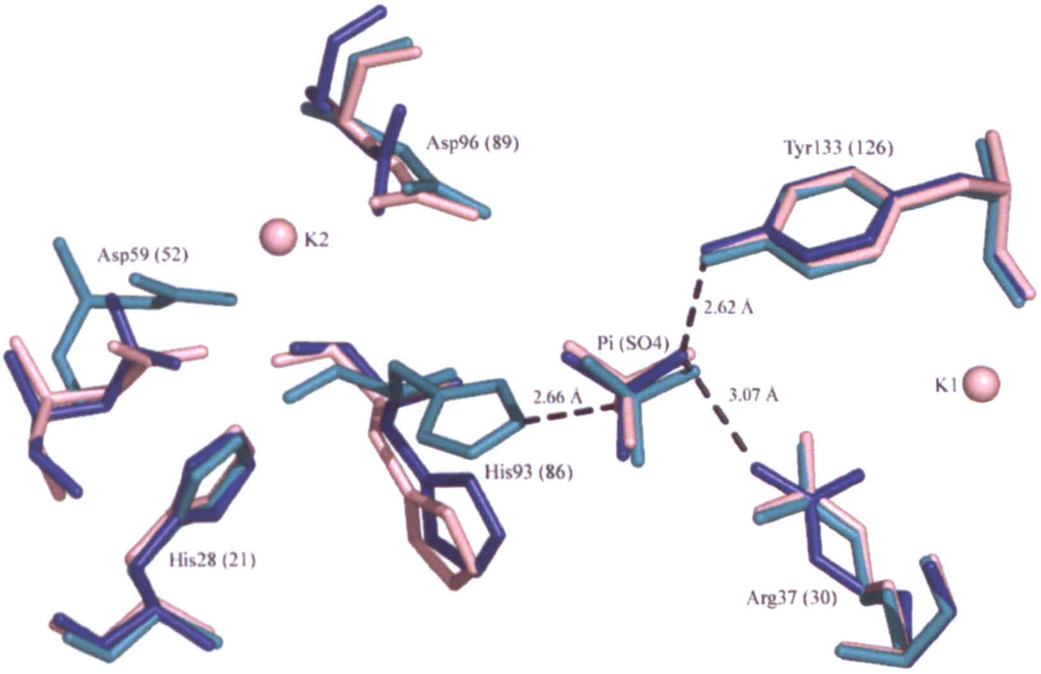


Figure 5.19: Superimposition of active site residues (chains A) with high rms deviations between the three Mtb-PPase structures. Mtb-PPase $P3_221$ (represented in dark blue); Mtb-PPase 1SXV (Tammenkoski *et al.*, 2005) (teal); and Mtb-PPase 1WCF (Benini and Wilson, to be published) (light pink). Mtb-PPase 1WCF potassium atoms are shown as spheres. Hydrogen bonds between Mtb-PPase $P3_221$ and the modelled phosphate, and between Mtb-PPase 1SXV and sulphate are shown as black broken lines.

5.8.5 Structural comparisons with type II inorganic pyrophosphatases

In comparison with the 34 X-ray crystal structures currently available for type I PPases, only six type II structures exist. These are from three organisms: *Bacillus subtilis* (Bs-PPase), complexed with Mn (Ahn *et al.*, 2001), SO_4, and $MnSO_4$ (both Fabrichniy *et al.*, 2004); *Streptococcus gordonii* (Sg-PPase), complexed with $MnSO_4$ (Ahn *et al.*, 2001) and $ZnSO_4$ (Fabrichniy *et al.*, 2004); and *Streptococcus mutans* (Sm-PPase), complexed with MgMnSO4 (Merckel *et al.*, 2001).

Superimposition of the main chain Cα atoms of the Mtb-PPase and Bs-PPase (SO_4) structures (chain A), gave an rms deviation of 15.89 Å, which is expected due to the low sequence identity between these two enzymes. As the primary sequences between type I and II PPases are so distinct (Fabrichniy *et al.*, 2004), sequence analysis was not performed. Bs-PPase forms a functional dimer, with each monomer folding into two distinct domains. The larger N-terminal region connects to the smaller C-terminal domain via a six-residue linker. This is in comparison with the biologically active Mtb-PPase hexamer, which forms from six, single domain monomers. The Mtb-PPase active site is partially buried within the enzyme, with most of the surface residues at the hexamer interface. In Bs-PPase, the active site resides within the N- and C-terminal domain interface (**figure 5.20**). Within the active site of both enzymes is the presence of two histidine residues, which bind to manganese in the Bs-PPase ($MnSO_4$) structure. The Mtb-PPase His93 is thought to be analogous to the Bs-PPase His98 (Tammenkoski *et al.*, 2005).

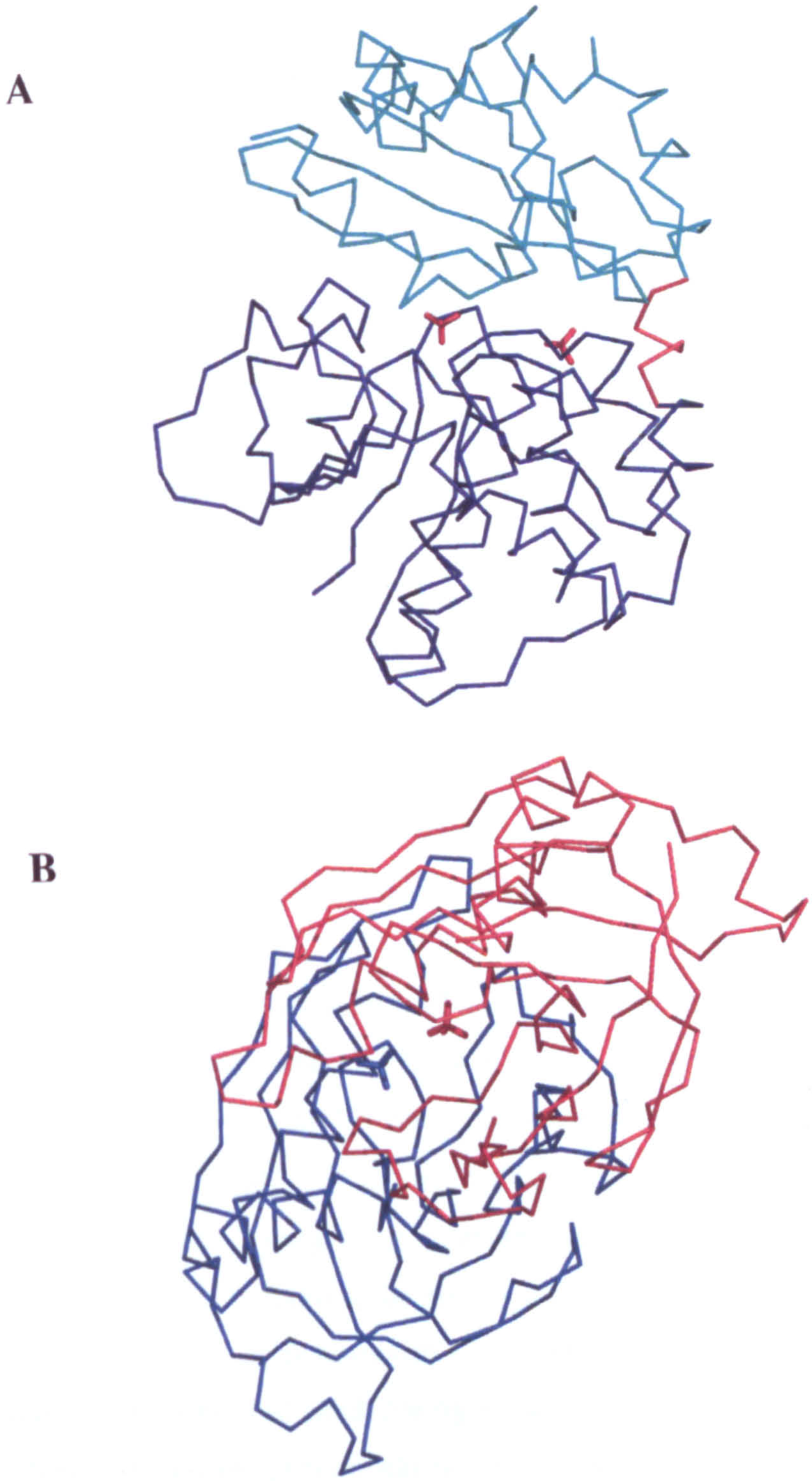


Figure 5.20: Comparison of the location of Bs-PPase and Mtb-PPase active sites. **(A)** Bs-PPase (SO_4) chain A: N-terminal domain shown in dark blue, C-terminal domain in cyan; and the linker region in deep pink (Fabrichniy *et al.*, 2004). The active site is located at the domain interface, represented by the two active site sulphate groups (deep pink sticks). **(B)** Mtb-PPase (PO_4): Chain C is shown as deep pink ribbons and the symmetrically-derived chain E, in blue. The active site is located at the hexameric interface, as represented by the active site phosphate groups, shown as sticks.

5.9 Conclusions

A type I inorganic pyrophosphatase from *Mycobacterium tuberculosis* has been over-expressed using a cell-free protein expression system, and its three dimensional structure has been determined to a resolution of 2.7 Å. Sequence analysis highlighted 17 residues explicitly conserved throughout type I PPases which form the active site, 13 of which participate in substrate/product or metal binding (Sivula *et al.*, 1999). Comparison of the structure with other prokaryotic type I PPases highlighted the striking similarity in the overall fold and orientation of active site residues. Although a structure for human PPase is not available, the already well documented similarity in the fold and conservation of active site residues for type I PPases suggest H-PPase may adopt analogous characteristics. This is further substantiated by a 52 % sequence identity between H-PPase and the well characterised Y-PPase. Whilst Y-PPase and Mtb-PPase share only 25 % identity, the core of the much larger Y-PPase monomer can be superimposed onto Mtb-PPase in a similar manner to that of prokaryotic PPases. Comparative analysis of the orientation of active site residues in Mtb-PPase and various ligand-bound homologues, emphasised the role of specific residues in phosphate and metal binding.

The higher resolution structure of Mtb-PPase in space group $P6_322$ (1SXV) identified two novel active site histidines, one of which, His93, coordinates with the modelled sulphate group (Tammenkoski *et al.*, 2005). Such coordination is not visible within the Mtb-PPase structure described here (pH 7.5), or in the $P6_322$ (pH 7.0) structure solved by Benini and Wilson (to be published), suggesting a pH-dependence. How crucial these histidines are for Mtb-PPase catalysis remains to be seen, since mutations of these residues only hamper activity in the presence of magnesium (Tammenkoski *et al.*, 2005). They propose that His93 may act as a general acid during catalysis in acidic conditions, however further investigations are required (Tammenkoski *et al.*, 2005).

Chapter 6 – Characterisation of cytochrome P450 125 (Rv3545c) from *Mycobacterium tuberculosis*

6.1 Introduction

An in-depth discussion of cytochrome P450 background has been described in section 1.3 and so will not be reiterated here. An interest in P450s has been developed during the course of this research, with initial work centring on the crystallisation of the plant CYP74C3, hydroperoxide lyase from *Medicago truncatula*. Despite considerable efforts in this area, no crystals suitable for X-ray diffraction were produced and so efforts were focused towards the expression and crystallisation of *Mycobacterium tuberculosis* targets. Identification of potential targets by literature review identified 22 probable P450s, one of which (Mtb-CYP125, encoded by the gene Rv3545c) was of particular interest due to its essential nature during *in vivo* infection in mice (Sassetti and Rubin, 2003) and the lack of existing structural information. This target was progressed into cell-free expression trials, however gave disappointing results (sections 4.2.4 to 4.2.5). Further attempts at obtaining soluble protein were successful using an *E. coli* expression system (sections 4.3.2 to 4.3.3), which allowed for progression into crystallisation trials and characterisation by spectroscopy, both of which are described in this chapter.

Further characterisation of Mtb-CYP125 by computational methods are also described in this chapter, to highlight sequence similarities between P450 homologues, and to predict secondary structural information and domain architecture.

All buffers described in this chapter are detailed in appendix 2.

6.2 Bioinformatics

6.2.1 Introduction

Bioinformatics can be described as the "mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information" (Fredj Tekaia at the Institut Pasteur). For this purpose, bioinformatics was used to compare homologues from the cytochrome P450 family, and to gain knowledge of the organisation of the Rv3545c gene.

6.2.2 Methods and results

6.2.2a Homologous protein searches

The Rv3545c protein sequence was obtained from the Tuberculosis Structural Genomics Consortium (TBSGC) and used to search for homologues within both the UniProt and the PDB databases using the NCBI BLAST2 blastp software (http://www.ebi.ac.uk/blastall/index.html).

The UniProt search provided confirmation that the correct sequence of the Rv3545c gene-product had been obtained, by showing a 100 % match with a putative cytochrome P450 125 from both *Mycobacterium tuberculosis* and *Mycobacterium bovis* (**table 6.1**). Two further bacterial Mtb-CYP125 homologues were identified in *Rhodococcus sp.* (RHA1) and *Nocardiodes sp.* (JS614), with sequence identities of 68 % and 55 % respectively. Additional searches using the NCBI database (http://www.ncbi.nlm.nih.gov) confirmed the existence of only four CYP125 proteins to date.

The Rv3545c gene product was also found to share 42 % sequence identity with *M. tb* CYP124, encoded by the gene Rv2266. P450 sequences from a further nine species were found with sequence identities above 35 %, from four *Mycobacterium* species, and one each from *Salinispora*, *Nocardia*, *Frankia*, *Rubrobacter*, and *Streptomyces*. Two *Mycobacterium* Linalool 8-monooxygenase sequences were identified, both with 37 % identity. Finally, a hypothetical protein from *Mycobacterium paratuberculosis* was found to share 82 % identity with Rv3545c, and a NigD from *Streptomyces violaceoniger* to share 35 %. See **table 6.1** for a summary of the output.

The PDB search identified eight unique sequences, whose protein structures are known, and which share 26 – 34 % identity with Rv3545c (**table 6.2**). The highest scoring of which was a P450terp from *Pseudomonas sp.* which shared 28 % sequence identity, demonstrating the low identity between P450s from different families.

Species	**Protein**	**Length**	**Sequence identify (%)**
Mycobacterium tuberculosis	Putative cytochrome P450 125	433	100
Mycobacterium bovis	Putative cytochrome P450 125	433	100
Mycobacterium paratuberculosis	Hypothetical protein	416	82
Mycobacterium vanbaalenii (PYR-1)	Cytochrome P450	419	74
Mycobacterium sp. (MCS)	Cytochrome P450	427	73
Mycobacterium sp. (KMS)	Cytochrome P450	427	73
Mycobacterium flavescens (PYR-GCK)	Cytochrome P450	417	74
Rhodococcus sp. (RHA1)	Cytochrome P450 125	471	68
Nocardia farcinica	Cytochrome P450 monooxygenase	422	68
Salinispora tropica (CNB-440)	Cytochrome P450	408	58
Streptomyces avermitilis	Cytochrome P450 hydroxylase	414	57
Nocardioides sp. (JS614)	Probable cytochrome P450 125	413	55
Mycobacterium tuberculosis	Putative cytochrome P450 124	428	42
Mycobacterium bovis	Putative cytochrome P450 124	428	42
Mycobacterium sp. (JLS)	Linalool 8-monooxygenase	433	37
Mycobacterium sp. (MCS)	Linalool 8-monooxygenase	433	37
Rubrobacter xylanophilus (DSM 9941 / NBRC 16129)	Cytochrome P450	414	36
Streptomyces violaceoniger	NigD	419	35
Frankia alni (ACN14a)	Putative cytochrome P450	423	36

Table 6.1: Output from an NCBI BLAST blastp UniProt database search of the Rv3545c gene product, Mtb-CYP125. Data taken from an output of the fifty highest scoring sequences. Sequences listed more than once were removed from the output.

Species	Protein	PDB ID	Score	Length	Sequence identity (%)
Pseudomonas sp.	Cytochrome P450terp	1CPT	420	428	28
Streptomyces venezuelae	Cytochrome P450pikC 107L1	2C7X	404	436	34
Saccharopolyspora erythraea	Cytochrome P450eryF 107A1	1Z8Q	384	404	30
Streptomyces coelicolor a3 (2)	Cytochrome P450 158A2	2D0E	369	407	31
Sulfolobus solfataricus	Cytochrome P450 CYP119	1IO8	357	368	28
Fusarium oxysporum	Cytochrome P450nor	1EHG	334	403	28
Citrobacter braakii	Cytochrome P450cin	1T2B	331	397	26
Amycolatopis orientalis	Cytochrome P450oxyB	1LGF	324	398	29

Table 6.2: Output from an NCBI BLAST blastp PDB database search of the Rv3545c gene product, Mtb-CYP125. Data taken from an output of the fifty highest scoring sequences. Sequences listed more than once were removed from the output.

6.2.2b Sequence alignment and secondary structure prediction

The protein sequence of Mtb-CYP125 was aligned with the homologous proteins identified in **table 6.2** using ClustalW (Thompson *et al.*, 1994 and www.ebi.ac.uk/clustalw/). Secondary structural elements of Rv3545c and the homologous P450terp (Hasemann *et al.*, 1994) were predicted using the PROF-sec function of PredictProtein (Rost *et al.*, 2003 and http://www.predictprotein.org). The aligned sequences were annotated to include secondary structural information of P450terp from both the prediction software and from crystallographic experiments, to allow a comparison of the two methods to be made (**figure 6.1**).

Due to the low sequence identity between P450s, the alignment cannot be considered complete and a number of key residues may not be marked as being conserved. Those which are, are generally located close to the haem-binding region, and in particular the Cys-loop (highlighted in **figure 6.1** with a thick underline). The only explicitly conserved P450 residue, the cysteinate proximal haem-ligand, is also conserved within Mtb-CYP125 at position Cys377. The three residues involved in hydrogen bonding between side chain nitrogen atoms and D-ring propionate oxygens in P450terp (Hasemann *et al.*, 1994): His124; Arg128; and His375, are also present. Four of the six residues found to form extended hydrogen networks with propionate-bound water molecules in P450terp are retained (Phe317, Arg318, Tyr341, and His375), however this region is generally less well conserved.

The EXXR motif (Glu305 and Arg308) within the K helix, conserved in most P450s, is also conserved in this protein, as is His375 from the "meander" region which forms hydrogen bonds with Glu305 homologues in P450BM-3 and P450cam (Peterson and Graham-Lorence, 1995). A region of the I helix, structurally conserved in P450BM-3 and P450cam, contains a residue essential for catalysis in some P450s, and is present in the Mtb-CYP125 sequence (Thr272), together with Glu271 which forms the (E/D)T pair described in sections 1.3.7a and 1.3.7e (Aikens and Sligar, 1994 and Tosha *et al.*, 2003).

PredictProtein predicted 13 discrete α-helical (~ 41 %) and 7 β-sheet (~ 8 %) regions, encoded by the Mtb-CYP125 amino acid sequence. A significant similarity was observed for the P450terp secondary structural information derived from both experimental and computational methods (**figure 6.1** and Haseman *et al.*, 1994). However the length of predicted secondary structure did not always match that of the actual data, and the computational method failed to recognise 3 α-helical and 7 β-sheet regions. Despite this, an approximation of Mtb-CYP125 secondary structure can be made using the computational data, when compared to that of P450terp. Similar secondary structural elements are apparent throughout the two proteins, of particular interest is the suggestion of an α-helix at Mtb-CYP125 residues Asp255 to Val267, a catalytically important region (αI) in a number of prokaryotic P450s with known structures (Poulos *et al.*, 1995). Surprisingly, no secondary structure was predicted at Mtb-CYP125 residues Lys101 to Val 111, which has been experimentally determined to be involved in substrate binding in P450terp, P450BM3, and P450cam (Hasemann *et al.*, 1995), however this was also missing from the P450terp prediction. Identical predictions of β-strand structure at Mtb-CYP125 residues Thr319 to Leu321 and corresponding P450terp residues were identified. This region has also been implicated in substrate binding.

```
                                                     αA'               αA         β1-1
Mtb-CYP125            ---------VSWNHQSVEIAVRRTTVPSPNLPPGFDFTDPAIYAERLPVAEFAELRSAAP 51
P450terp (1CPT 28 %)  ---------------------MDARATIPEHIARTVILPQGYADDEVIYPAFKWLRDEQP 39
P450pikC (2C7X 34 %)  MGSSHHHHHHSSGLVPRGSHMRRTQQGTTASPPVLDLGALGQDFAADPYPTYARLRAEGP 60
P450eryF (1Z8O 30 %)  ------------------------------MTTVPDLES--DSFHVDWYRTYAELRETAP 28
CYP158A2 (2D0E 31 %)  ----------------------MTEETISQAVPPVRDWPAVDLPGSDFDPVLTELMREGP 38
CYP119 (1IO8 28 %)    -----------------------------------------------MYDWFSEMRKKDP 13
P450nor (1EHG 28 %)   ------------------------------MASGAPSFPFSRASGPEPPAEFAKLRATNP 30
P450cin (1T2B 26 %)   ----------------------------TSLFTTADHYHTPLGPDGTPHAFFEALRDEAE 32
P450oxyB (1LGF 29 %)  ----------------------MSED----------DPRPLHIRRQGLDP-ADELLAAGA 27

                                         β1-2    αβ       β1-5             αβ'
Mtb-CYP125            IWWNGQDPGKGGGFHDGGFWAITKLNDVKEISRHSDVFSSYENGVIPRFKNDIAREDIEV 111
P450terp              LAMAHIEGYDP-------MWIATKHADVMQIGKQPGLFSNAEGSEILYDQNNEAFMRSIS 92
P450pikC              AHRVRTPEGDE-------VWLVVGYDRARAVLADPRFSKDWRNSTTPLTEAEAALN---- 109
P450eryF              VTPVR-FLGQD-------AWLVTGYDEAKAALSDLRLSSDPKKKYPGVEVEFPAYLGFPE 80
CYP158A2              VTRISLPNGE------GWAWLVTRHDDVRLVTND-PRFGREAVMDRQVTRLAPHFIPARG 91
CYP119                VYYDG------------NIWQVLSYRYTKEVLNNFSKFSSDLTGYHERLEDLRNGKIRFD 61
P450nor               VSQVKLFDGS-------LAWLVTKHKDVCFVATSEKLSKVRTRQGFPELSASGKQAAKAK 83
P450cin               TTPIGWSEAYGG------HWVVAGYKEIQAVIQNTKAFSNKGVTFPRYETGEFELMMAGQ 86
P450oxyB              LTRVTIGSGADA----ETHWMATAHAVVRQVMGDHQQFSTRRRWDPRDEIGGKGIFRPRE 83

                                          αC         αC'          αD         β3-1
Mtb-CYP125            QR-----FVMLNMDAPHHTRLRKIISRGFTPRAVGRLHDELQERAQKIAAEAAAAG---- 162
P450terp              GGCPHVIDSLTSMDPPTHTAYRGLTLNWFQPASIRKLEENIRRIAQASVQRLLDFDG--- 149
P450pikC              -------HNMLESDPPRHTRLRKLVAREFTMRRVELLRPRVQEIVDGLVDAMLAAPD--G 160
P450eryF              DVRNYFATNMGTSDPPTHTRLRKLVSQEFTVRRVEAMRPRVEQITAELLDEVGDS----G 136
CYP158A2              -----AVG---FLDPPDHTRLRRSVAAAFTARGVERVRERSRGMLDELVDAMLRAG---P 140
CYP119                IP---TRYTMLTSDPPLHDELRSMSADIFSPQKLQTLETFIRETTRSLLDSIDPRE---- 114
P450nor               P-------TFVDMDPPEHMHQRSMVEPTFTPEAVKNLQPYIQRTVDDLLEQMKQKGCANG 136
P450cin               -------------DDPVHKKYRQLVAKPFSPEATDLFTEQLRQSTNDLIDARIELG---- 129
P450oxyB              -----LVGNLMDYDPPEHTRLRRKLTPGFTLRKMQRMAPYIEQIVNDRLDEMERAG---S 135

                          αE'       αE              αF                           αG
Mtb-CYP125            SGDFVEQVSCELPLQAIAGLLGVPQEDRGKLFHWSNEMTGNEDPEYAHIDP--------- 213
P450terp              ECDFMTDCALYYPLHVVMTALGVPEDDEPLMLKLTQDFFGVHEPDEQAVAAPRQSADEAA 209
P450pikC              RADLMESLAWPLPITVISELLGVPEPDRAAFRVWTDAFVFPDD--PAQAQ---------- 208
P450eryF              VVDIVDRFAHPLPIKVICELLGVDEKYRGEFGRWSSEILVMDPERAEQRG---------- 186
CYP158A2              PADLTEAVLSPFPIAVICELMGVPATDRHSMHTWTQLILSSSHG-AEVSE---------- 189
CYP119                -DDIVKKLAVPLPIIVISKILGLPIEDKEKFKEWSDLVAFRLGKPGEIFEL--------- 164
P450nor               PVDLVKEFALPVPSYIIYTLLGVPFND---LEYLTQQNAIRTNGSSTAREA--------- 184
P450cin               EGDAATWLANEIPARLTAILLGLPPEDGDTYRRWVWAITHVENPEEGAEIF--------- 180
P450oxyB              PADLIAFVADKVPGAVLCELVGVPRDDRDMFMKLCHGHLDASLS-QKRRA---------- 184

                                                     αH  β5-1  β5-2     αI
Mtb-CYP125            ---KASSAELIGYAMKMAEEKAKNPADDIVTQLIQADID-GEKLSDDEFGFFVVMLAVAG 269
P450terp              RRFHETIATFYDYFNGFTVDRRSCPKDDVMSLLANSKLD-GNYIDDKYINAYYVAIATAG 268
P450pikC              ----TAMAEMSGYLSRLIDSKRGQDGEDLLSALVRTSDEDGSRLTSEELLGMAHILLVAG 264
P450eryF              ----QAAREVVNFILDLVERRRTEPGDDLLSALIRVQDDDDGRLSADELTSIALVLLLAG 242
CYP158A2              ----RAKNEMNAYFSDLIGLRSDSAGEDVTSLLGAAVGR--DEITLSEAVGLAVLLQIGG 243
CYP119                ---GKKYLELIGYVKDHLN-----SGTEVVSRVVNSNLS------DIEKLGYIILLLIAG 210
P450nor               ---SAANQELLDYLAILVEQRLVEPKDDIISKLCTEQVK-PGNIDKSDAVQIAFLLLVAG 240
P450cin               -------AELVAHARTLIAERRTNPGNDIMSRVIMSKID-GESLSEDDLIGFFTILLLGG 232
P450oxyB              ----ALGDKFSRYLLAMIARERKEPGEGMIGAVVAEYG---DDATDEELRGFCVQVMLAG 237

                                            αJ            αK            β1-4    β2-1
Mtb-CYP125            NETTRNSITQGMMAFAEHPDQWELYKKVRP--ETAADEIVRWATP--VTAFQRTALRDYE 325
P450terp              HDTTSSSSGGAIIGLSRNPEQLALAKSDPALIPRLVDEAVRWTAP--VKSFMRTALADTE 326
P450pikC              HETTVNLIANGMYALLSHPDQLAALRADMTLLDGAVEEMLRYEGP-VESATYRFPVEPVD 323
P450eryF              FEASVSLIGIGTYLLLTHPDQLALVRRDPSALPNAVEEILRYIAP-PETTT-RFAAEEVE 300
CYP158A2              -EAVTNNSGQMFHLLLSRPELAERLRSEPEIRPRAIDELLRWIPHRNAVGLSRIALEDVE 302
CYP119                NETTTNLISNSVIDFTRFN-LWQRIR-EENLYLKAIEEALRYSPP--VMRTVRKTKERVK 266
P450nor               NATMVNMIALGVATLAQHPDQLAQLKANPSLAPQFVEELCRYHTA-VALAIKRTAKEDVM 299
P450cin               IDNTARFLSSVFWRLAWDIELRRRLIAHPELIPNAVDELLRFYGP---AMVGRLVTQEVT 289
P450oxyB              DDNISGMIGLGVLAMLRHPEQIDAFRGDEQSAQRAVDELIRYLTVPYSP-TPRIAREDLT 296
```

Continued on page 197.

```
                 β2-2     β1-3  αK'                                  αL
Mtb-CYP125   LSGVQIKKGQRVVMFYRSANFDEEVFQDPFTFNILR--NPNPHVGFGGTGAHYCIGANLA 383
P450terp     VRGQNIKRGDRIMLSYPSANRDEEVFSNPDEFDITR--FPNRHLGFG-WGAHMCLGQHLA 383
P450pikC     LDGTVIPAGDTVLVVLADAHRTPERFPDPHRFDIRR--DTAGHLAFG-HGIHFCIGAPLA 380
P450eryF     IGGVAIPQYSTVLVANGAANRDPKQFPDPHRFDVTR--DTRGHLSFG-QGIHFCMGRPLA 357
CYP158A2     IKGVRIRAGDAVYVSYLAANRDPEVFPDPDRIDFER--SPNPHVSFG-FGPHYCPGGMLA 359
CYP119       LGDQTIEEGEYVRVWIASANRDEEVFHDGEKFIPDR--NPNPHLSFG-SGIHLCLGAPLA 323
P450nor      IGDKLVRANEGIIASNQSANRDEEVFENPDEFNMNRKWPPQDPLGFG-FGDHRCIAEHLA 358
P450cin      VGDITMKPGQTAMLWFPIASRDRSAFDSPDNIVIER--TPNRHLSLG-HGIHRCLGAHLI 346
P450oxyB     LAGQEIKKGDSVICSLPAANRDPALAPDVDRLDVTR--EPIPHVAFG-HGVHHCLGAALA 353

                           β3-3      β4-1        β4-2  β3-2
Mtb-CYP125   RMTINLIFNAVADHMPDLKPISAP---ERLRSGWLNGIKHWQVDYTGRCPVAH---      433
P450terp     KLEMKIFFEELLPKLKSVELSGPPR---LVATNFVGGPKNVPIRFTKA--------      428
P450pikC     RLEARIAVRALLERCPDLALDVSPGELVWYPNPMIRGLKALPIRWRRGREAGRRTG      436
P450eryF     KLEGEVALRALFGRFPALSLGIDADDVVWRRSLLLRGIDHLPVRLDG---------      404
CYP158A2     RLESELLVDAVLDRVPGLKLAVAPEDVPFKKGALIRGPEAL---------------      400
CYP119       RLEARIAIEEFSKRFRHIEILDTEK----VPNEVLNGYKRLVVRLKSNE-------      368
P450nor      KAELTTVFSTLYQKFPDLKVAVPLGKINYTPLNRDVGIVDLPVIF-----------      403
P450cin      RVEARVAITEFLKRIPEFSLDPNKE--CEWLMGQVAGMLHVPIIFPKGKRLSE---      397
P450oxyB     RLELRTVFTELWRRFPALRLADPAQDTEFRLTTPAYGLTELMVAW-----------      398
```

Figure 6.1: Multiple sequence alignment of Mtb-CYP125 homologues, aligned using ClustalW (Thompson *et al*, 1994). Sequence identities to Mtb-CYP125 are shown in parentheses. Key residues are highlighted: conserved residues (boldface); explicitly conserved proximal cysteinate ligand (red); generally conserved Cys-loop region (thick underline). Underlined residues and annotations correspond to the secondary structure of P450terp as determined by X-ray crystallography (Hasemann *et al.*, 1994). Coloured text represents predicted secondary structural elements (α-helices in red and β-sheets in blue) for Mtb-CYP125 and P450terp, as determined by sequence analysis using the PROF-sec function of PredictProtein (Rost *et al.*, 2003).

6.2.2c Domain organisation

The SMART database (genomic mode) was used to characterise the domain architecture of the Rv3545c gene product. PFAM domain and signal peptide searches performed on the Mtb-CYP125 protein sequence identified only one domain, that of a cytochrome P450. This is to be expected as despite requiring a redox partner, only a very few P450s encode this on a single polypeptide (**figure 6.2**), as described in section 1.3.2.

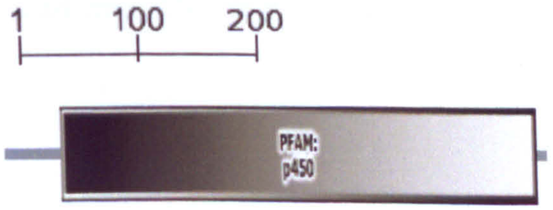


Figure 6.2: Domain organisation of Mtb-CYP125, taken from the SMART database (http://smart.embl-heidelberg.de/).

6.3 Ultra-violet/visible absorption spectroscopy

6.3.1 Introduction

Spectroscopic techniques are widely used during the characterisation of haem proteins and provide abundant information about the molecule's identity, purity, and specific activity. Such methods are also useful in determining oxidation state, estimating spin state, and assessing ligand binding (Li, 2001). Spectral measurements of cytochrome P450s are especially well established and a number of experiments are discussed further here.

Due to the highly characteristic shift in maximal absorbance to 450 nm upon binding of CO to ferrous P450 (Omura and Sato, 1964), the presence of haem-thiolate proteins can be identified by spectroscopy. Binding of substrate ligands can often be inferred from spectral changes which occur, due to the accompanying spin-state shifts, induced by the blocking of water access to the distal haem iron position (Sligar, 1976 and Li, 2001). It may also be possible to identify the oxidation states of a P450 haem iron, although spectral shifts between the two states are often negligible (Li, 2001).

6.3.2 Methods and results

Mtb-CYP125 was expressed and purified as described in section 4.3.2e and concentrated to 40 mg/ml (826 μm) using a Vivapsin-6 30kDa MWCO centrifuge filtration unit. Concentration was determined by absorbance at 280 nm, using a theoretical extinction coefficient of 62.8 $M^{-1}cm^{-1}$ (TBSGC). All spectral measurements were performed using UVWinLab 2 software on a Lambda16 dual UV/visible spectrophotometer (both Perkin Elmer). Scans were measured from 270 nm to 700 nm with a 1 nm slit width.

6.3.2a Spin state and substrate binding

Initial Measurements

A spectrum was measured after each purification step. Two identical 500 µl quartz cuvettes containing the appropriate purification buffer (**table 6.3**) were used to blank the system. The buffer in one cuvette was replaced with Mtb-CYP125, diluted to 5 µm in the same buffer, and a spectral scan was measured.

Mtb-CYP125 purification stage	Buffer	Buffer composition
Nickel-affinity chromatography	125-NiC	500 mM KPi pH 7.4 300 mM Imidazole 10 mM β-mercaptoethanol
Gel filtration chromatography	125-GF	50 mM KPi pH 7.4 500 mM KCl 1 mM DTT

Table 6.3: Spectral characterisation of Mtb-CYP125 after each purification step. Buffers shown were used to dilute protein stocks and as blanks.

The nickel-affinity preparation gave a spectrum characteristic of oxidised P450 in a low-spin (LS) state (Li, 2001), as defined by a sharp 426 nm Soret peak with fainter α and β peaks at 540 nm and 575 nm (**figure 6.3**). After further purification by gel filtration, a blue type-I shift typical of a high-spin (HS) system was observed (Li, 2001). The Soret peak shifted to 392 nm and a small, broader peak at 640 nm, replaced the previous α and β peaks (**figure 6.3**). A similar shift is also seen in the plant P450, hydroperoxide lyase (HPL), from *Medicago truncatula* (Hughes *et al.*, 2006), and in P450BM-3 (Li, 2001) (see section 1.3.8). This is thought to be due to a water molecule coordinating to the distal haem-iron position in the LS system, which is sterically blocked by the presence of "substrate" in the HS system.

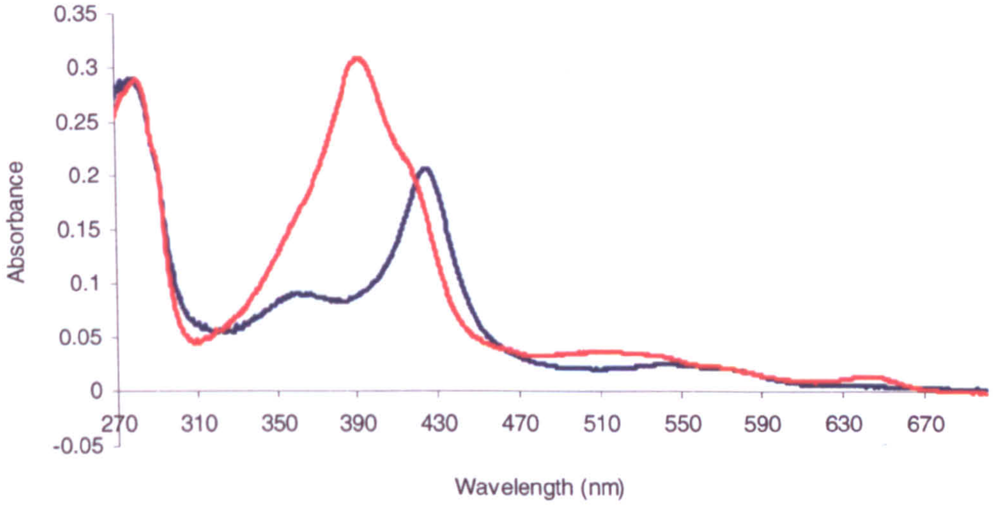


Figure 6.3: UV/visible spectra of Mtb-CYP125 after purification by (1) nickel-affinity chromatography in LS form (blue) and (2) nickel-affinity and gel filtration chromatography, in HS form (red). Curves normalised at 280 nm.

Inhibition by imidazole

It was hypothesised that the different buffer components (**table 6.3**) used during each purification step may account for the spectral differences observed. As described in section 1.3.6a, imidazole acts as an inhibitor of P450s by binding reversibly to the distal site of the haem-iron, causing a type-II red shift in the Soret maximum (section 1.3.8).

This would account for the LS state of the nickel-affinity preparation in the presence of 300 mM imidazole. Desalting by gel filtration would, in the absence of substrate, result in the transfer of water back to the sixth position. However in the presence of substrate, water access is blocked and the haem shifts to a HS fifth co-ordinated system.

A spectral scan of the nickel-affinity preparation (5 µm) was performed using 125-NiC as a blank (scan 1, **figure 6.4**). A further 5 µm was diluted into imidazole-free buffer (125-Ni-I) and the spectrum was repeated, using 125-Ni-I buffer as a blank (scan 2). Finally, the Mtb-CYP125 used in scan 2 was diluted to 2.5 µm in imidazole buffer (end concentration 300 mM) and the spectrum repeated using 125-NiC as a blank (scan 3).

Removing the imidazole from Mtb-CYP125 resulted in a LS to HS shift, characterised by the change in Soret peak from 426 nm to 392 nm. A return to the LS system was accomplished by reintroducing imidazole into the system, demonstrating the reversible inhibition of Mtb-CYP125 by this compound (**figure 6.4**).

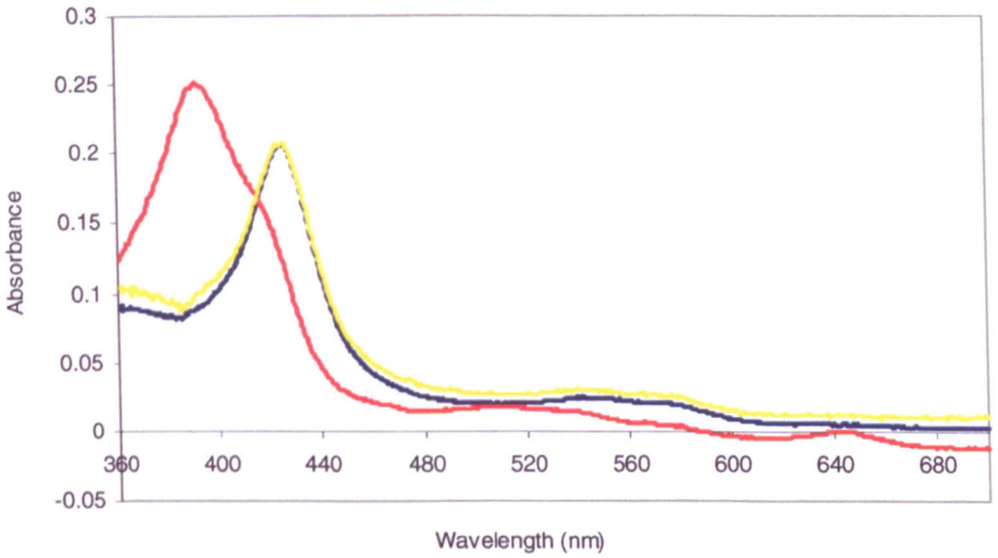


Figure 6.4: UV/visible spectra of Mtb-CYP125: Inhibition by imidazole. Mtb-CYP125 in: (1) 300 mM imidazole buffer, in LS form (blue); (2) diluted into imidazole-free buffer, in HS form (red); and (3) 300 mM imidzole reintroduced, in LS form (yellow). All curves normalised at 280 nm.

Identification of substrate

The ability of the haem iron to alternate between two spin states suggests that a (pseudo) substrate must remain in the system when not bound. As phosphate was included in both purification buffers at high concentrations (above 500 mM), it was thought that it might be mimicking a natural substrate, blocking access of water to the distal haem iron position.

A spectral scan of the gel filtration preparation (5 µm) was performed using 125-GF as a blank. The same sample was then buffer exchanged into a phosphate-free buffer (buffer 125-P) using a Vivaspin-6 30 kDa MWCO filtration unit. 125-P buffer was used to blank the spectrophotometer. No return to a low-spin system was observed upon dilution of phosphate from the system (**figure 6.5**).

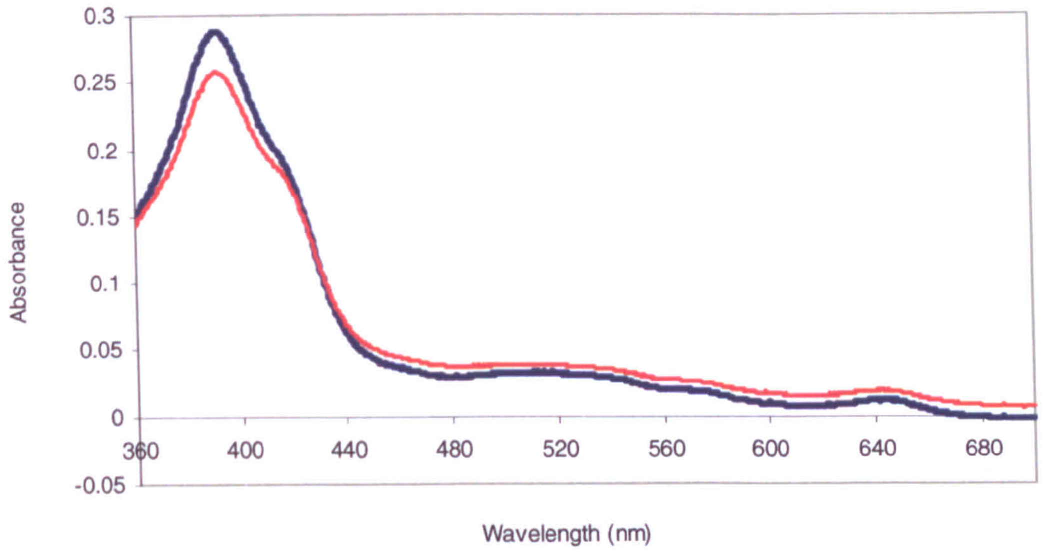


Figure 6.5: UV/visible spectra of Mtb-CYP125: Identification of substrate. Mtb-CYP125 in: (1) 500 mM phosphate buffer, in HS form (blue); and (2) buffer exchanged into phosphate-free buffer, in HS form (red). Curves normalised at 280 nm.

Cation binding site

As described in section 1.3.7h, structural characterisation of P450cam identified an increase in substrate affinity upon cation binding (Peterson, 1971). Although no such binding site has been identified in any other structurally characterised P450 (Li, 2001), it was thought that the high concentrations (above 500 mM) of potassium included in the purification buffers may be contributing to "substrate" binding affinity. Also, increased salt concentrations have been found to induce a spin-shift (low to high-spin) in a P450cam (Lange *et al.*, 1980), CYP1A2 (Yun *et al.*, 1996), and CYP2B1 (Yun *et al.*, 1998).

A spectral scan of the gel filtration preparation (5µm) was performed using 125-GF as a blank. A further 5 µm of Mtb-CYP125 was buffer exchange into a potassium/phosphate-free buffer (125-K) and a spectrum was obtained using 125-K as a blank. No shift occurred upon dilution of potassium and phosphate from the gel filtration preparation of Mtb-CYP125 (**figure 6.6**).

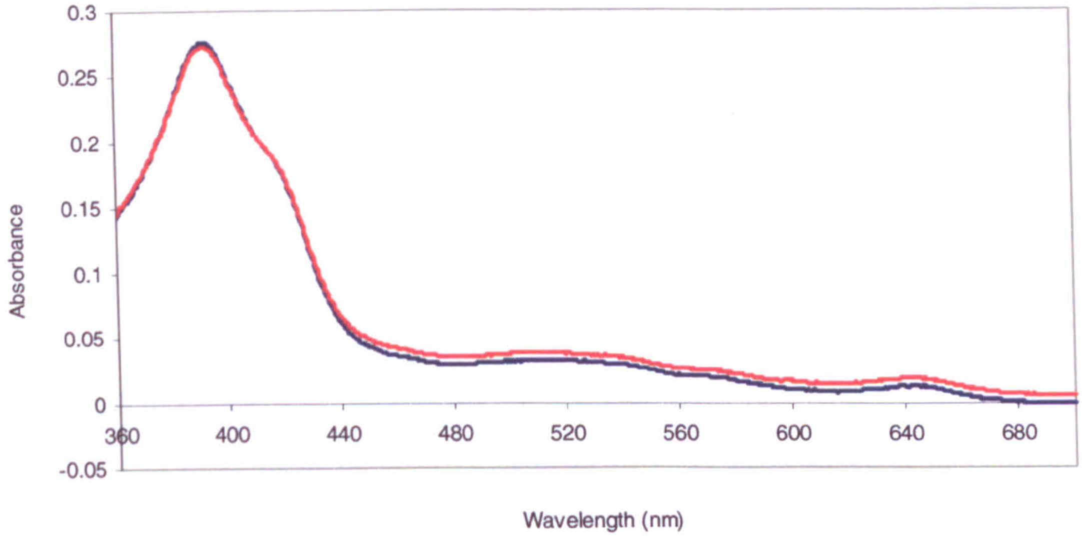


Figure 6.6: UV/visible spectra of Mtb-CYP125: Cation binding site. Mtb-CYP125 in: (1) 50 mM KPi, pH 7.4 + 500 mM KCl, in HS form (blue); (2) buffer exchanged into potassium and phosphate-free buffer, also in HS form (red). Curves normalised at 280 nm.

To see if removal of the two buffer components from the nickel-affinity preparation prevented a HS shift upon dilution of imidazole, Mtb-CYP125 (5 μm) was buffer exchanged into potassium/phosphate/imidazole-free buffer (125-K) and a spectrum recorded. Again no spin shift occurred, signifying either strong "substrate" binding or that potassium exerts little or no effect on Mtb-CYP125 substrate binding (**figure 6.7**).

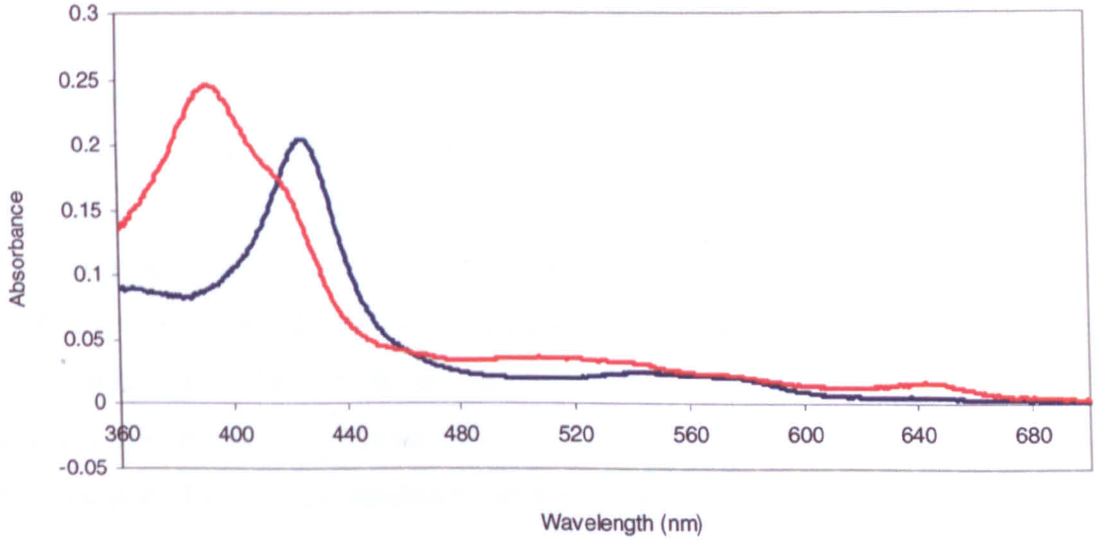


Figure 6.7: UV/visible spectra of Mtb-CYP125: Cation binding site (in the presence of imidazole). Mtb-CYP125 in: (1) 500 mM KPi + 300 mM Imidazole, in LS form (blue); (2) buffer exchanged into potassium/phosphate/imidazole-free buffer, in HS form (red). Curves normalised at 280 nm.

6.3.2b CO-binding assay

To confirm the function of the Rv3545c gene product, inferred through sequence homology to be a cytochrome P450, the CO-binding assay was performed (see section 1.3.8). The protocol was modified from Nelson (1998), in that dithionite was used to reduce the protein prior to inclusion of CO gas.

Mtb-CYP125 in high-spin state (substrate-bound)

Briefly, two identical 500 µl quartz cuvettes were sealed with rubber caps and degassed with N_2. 125-GF buffer (without DTT) was degassed in the same way and used to dilute Mtb-CYP125 to 10 µm in a sealed, degassed flask. A 50 mM sodium dithionite (Sigma) solution was made fresh before each experiment in the degassed buffer. Both cuvettes were blanked in 500 µl of degassed buffer over a range of 240 to 700 nm. 260 µl of buffer was removed from the "sample" cuvette and replaced with 250 µl of HS, substrate-bound Mtb-CYP125 (5 µm), and 10 µl of the dithionite solution (1 mM). The cuvette was incubated for ten minutes at room temperature, to allow reduction of the Mtb-CYP125 haem iron. The cuvette was then gently bubbled with 5 ml of CO gas, using a syringe. A "blank" spectrum was also recorded, identical to that of the "sample" cuvette, except lacking CO. Difference spectra were obtained by subtracting the "blank" from that of the "sample" data.

The spectrum produced from oxidised HS Mtb-CYP125 had a Soret peak at 392 nm with a 418 nm shoulder (**figure 6.8**), which became much broader upon reduction with dithionite. The typical Soret peak of ferrous P450s is around 408 nm, so this suggests reduction had at least partially occurred (Li, 2001). Incubating the reduced sample with CO for ten minutes resulted in the characteristic 450 nm absorbance maximum, together with a distinct 420 nm peak of similar size. The difference spectrum offers a more accurate representation of this data, as the contribution of reduced Mtb-CYP125 alone is removed from the 420 nm peak in the presence of CO. The resultant spectrum illustrates a wide trough at about 392 nm with a peak at 450 nm, signifying conversion of Mtb-CYP125 to the Mtb-CYP125-CO complex. The peak seen at 420 nm may represent partial conversion of Mtb-CYP125 to the inactive cytochrome P420, the value of which is negative due to the broadening of the resting P450 Soret peak upon reduction, resulting in a greater absorbance at 420 nm than that of the CO-complex.

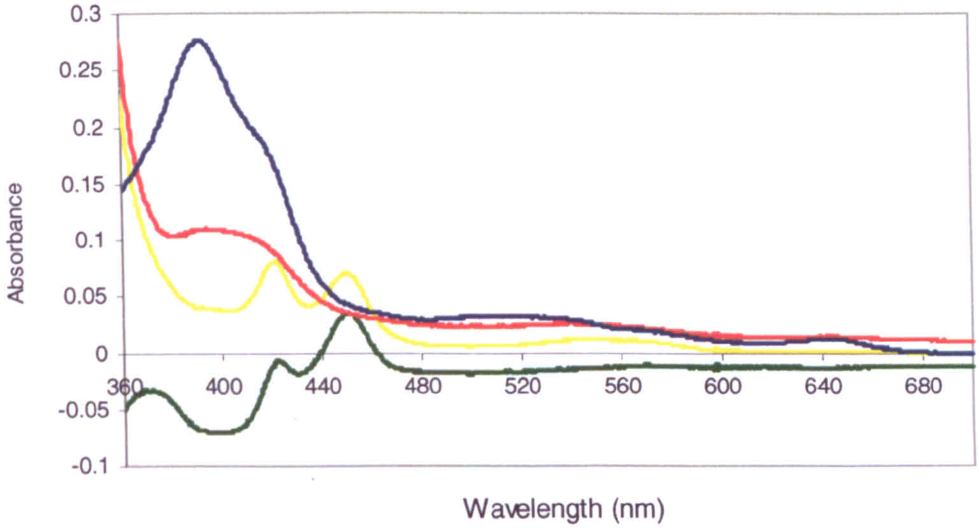


Figure 6.8: UV/visible spectra of Mtb-CYP125 (substrate-bound, HS): CO-binding assay. Mtb-CYP125 in: (1) ferric state (blue); (2) reduced by 1 mM sodium dithionite ("blank") (red); (3) ferrous state, bubbled with 5 ml of CO ("sample") (yellow); (4) the difference spectrum obtained by subtracting the "blank" from the "sample" data (green).

Effect of glycerol on P450 conversion to P420 (substrate-bound)

Due to the partial conversion of ferrous P450 to its inactive form, P420, in the presence of CO, an additional preparation of Mtb-CYP125 was purified in buffers containing glycerol. Glycerol is known to protect the hydrophobic region within P450s, thereby limiting the damage caused to the proximal cysteinate ligand (Falzon *et al.*, 1986 and Nebbia *et al.*, 1999).

1L of Mtb-CYP125 culture was grown and purified as described in section 4.3.2e, with the exception that all buffers contained 20 % glycerol. The protein was concentrated to 826 µm in 125-GF buffer including 20 % glycerol. Both preparations of Mtb-CYP125 (with and without glycerol) were subjected to the CO-binding assay, as described previously. **Figure 6.9** shows the difference spectra for both preparations.

Some protection was conferred by including 20 % glycerol in the protein preparation, evidenced by a larger P450 to P420 ratio than for the glycerol-free protein, however P420 conversion was not eliminated entirely.

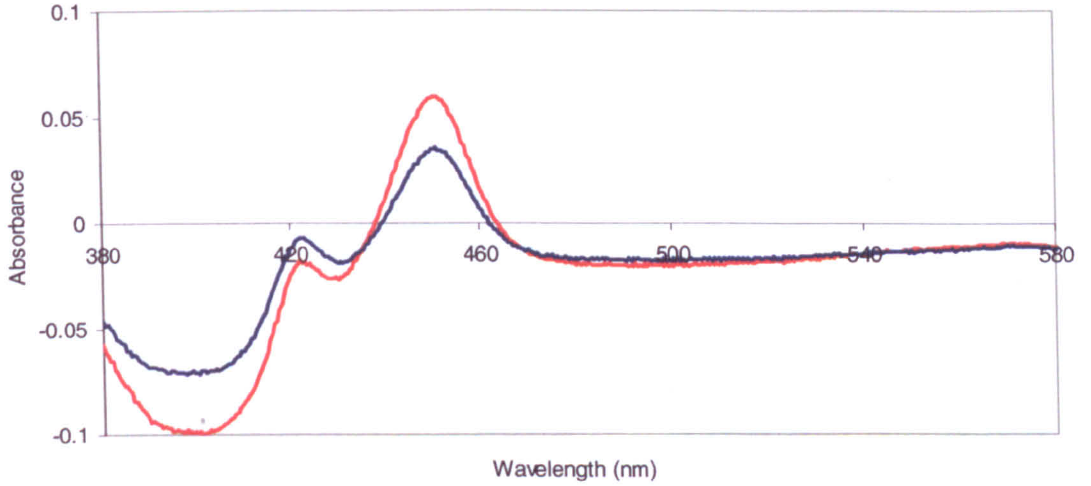


Figure 6.9: UV/visible spectra of ferrous Mtb-CYP125 (substrate-bound, HS) in the presence of CO: Effect of glycerol on P450 conversion to P420. Difference spectra of Mtb-CYP125 in: (1) buffer 125-GF (blue) and (2) buffer 125-GF + 20 % glycerol (red).

Effect of time on inactivation of P450 to P420 (substrate-bound)

To investigate the conversion of P450 to P420 over time, CO-difference spectra were recorded at increasing intervals after reduction. 5 ml of CO was bubbled anaerobically into the ferric Mtb-CYP125 cuvette, before adding 1 mM sodium dithionite. Spectra were recorded over 0 to 25 minutes. Identical 'blank' spectra, lacking CO, were also recorded.

Figure 6.10 shows a time-dependent increase in both P420 and P450 species, with no obvious conversion of the P450-CO complex to its inactive form.

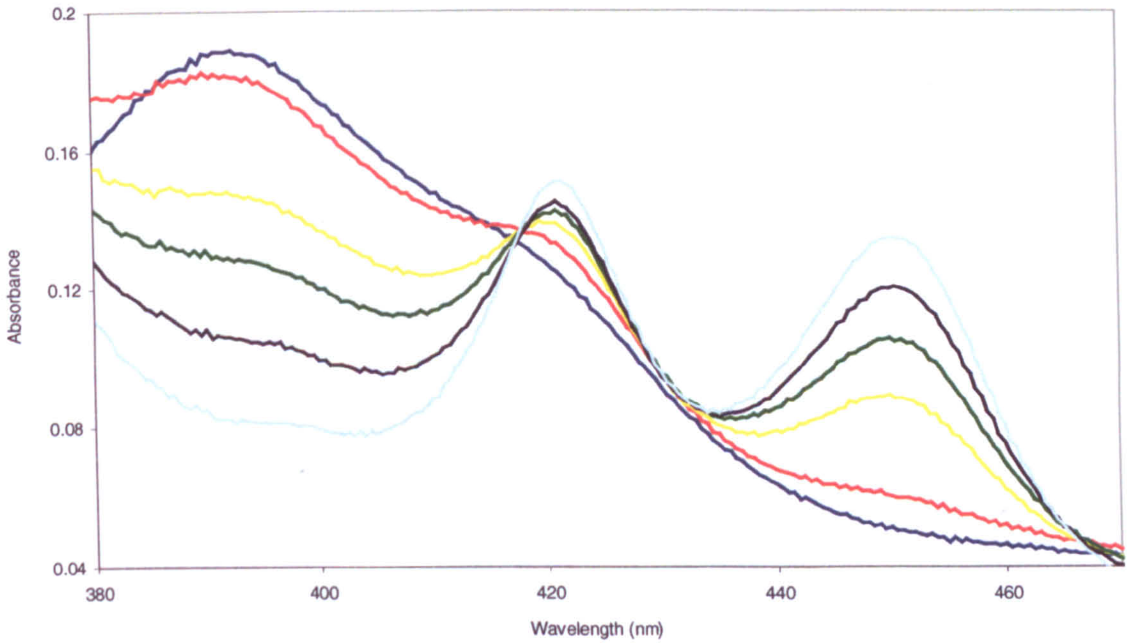


Figure 6.10: UV/visible spectra of ferrous Mtb-CYP125 (substrate-bound, HS) in the presence of CO: Effect of time on conversion of P450 to P420. Spectra of Mtb-CYP125 in buffer 125-GF, bubbled with 5 ml of CO. Time after reduction by 1 mM dithionite: 0 min (blue); 4 mins (red); 10 mins (yellow); 15 mins (green); 20 mins (black); and 25 mins (light blue).

Mtb-CYP125 in low-spin state (substrate-free)

The CO-binding assay was performed, following nickel-affinity chromatography, using Mtb-CYP125 in a substrate-free, low-spin state. Buffer 125-NiC (without β-mercaptoethanol) was used as a blank. The redox potential of substrate-free P450 differs from that of the substrate complex, as observed by a shift from -170 to -300 mV in P450cam upon loss of substrate (see section 1.3.5) (Sligar and Gunsalus, 1976 and Li, 2001). In addition to this, inhibitors such as imidazole which bind to the distal haem iron position, also induce an increase in redox potential (Ortiz de Montellano and Correia, 1995). As the redox potential of dithionite is - 550 mV, its ability to reduce the LS imidazole-bound Mtb-CYP125 is decreased, in comparison with the HS form. To counteract this, 10 mM of dithionite was used to reduce the protein and a spectrum was recorded over 0 to 60 minutes in the presence of CO, to ensure full reduction of the P450.

The spectrum produced from oxidised LS Mtb-CYP125 exhibits a well-defined Soret peak at 426 nm (**figure 6.11**), with a very slight blue shift upon reduction. Incubating the reduced sample with CO for 25 minutes resulted in a further blue shift to about 420 nm, together with the characteristic 450 nm absorbance maximum. The difference spectrum clearly highlights the 426 nm trough together with the 450nm Soret peak, along with a negative 420 nm peak.

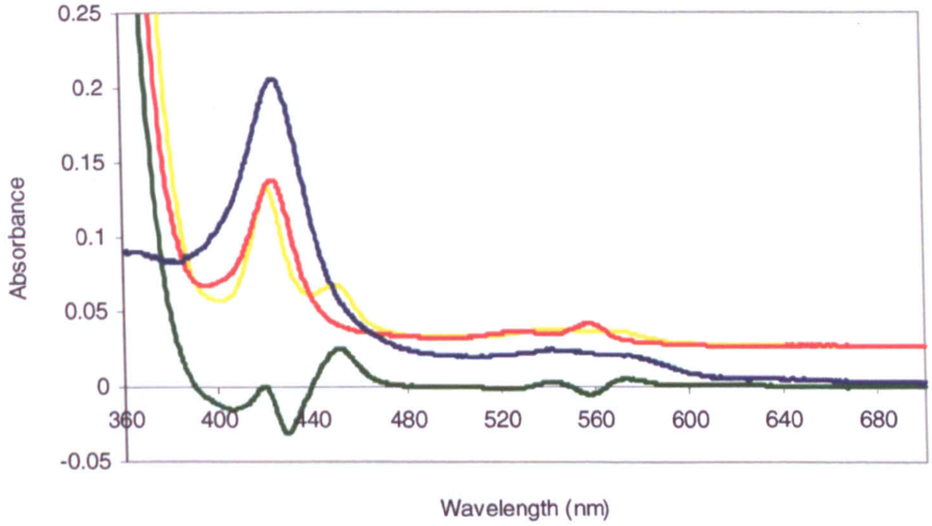


Figure 6.11: UV/visible spectra of Mtb-CYP125 (substrate-free, LS): CO-binding assay. Mtb-CYP125 in: (1) ferric state (blue); (2) reduced with 10 mM sodium dithionite ("blank") (red); (3) ferrous state, bubbled with 5 ml of CO ("sample") (yellow); and (4) the difference spectrum obtained by subtracting the "blank" from the "sample" data (green).

6.4 Circular dichroism

6.4.1 Introduction

In order to identify secondary structural elements within Mtb-CYP125, and to complement data obtained through computational methods (section 6.2.2b), circular dichroism (CD) was performed. To enable comparisons to be made with another P450, a CD spectrum of hydroperoxide lyase (CYP74C3) from *Medicago truncatula* (HPL) was also recorded.

6.4.2 Methods and results

All CD data were collected on station 12.1 at the Synchrotron Radiation Source, Daresbury. The detector was first calibrated using a standard protein solution of cunphosulphonic acid (CSA) at a concentration of 10 mg/ml. A blank spectrum of water alone was subtracted from the CSA data. This resulted in a 2 : 1 ratio between a negative peak at 192 nm and positive peak at 290 nm, characteristic of a successful calibration. Prior to sample application, the system was purged with nitrogen gas to prevent absorption of contaminating oxygen during data collection.

Purified HPL was kindly provided by Drs. Richard Hughes and Eric Belfield at the John Innes Centre, Norwich. Protein concentrations were determined by absorbance at 280 nm using theoretical extinction coefficients derived from ProtParam (http://www.expasy.org/tools/protparam.html). Mtb-CYP125 and HPL were diluted to 5 mg/ml in low salt buffer: Mtb-CYP125 in 50 mM potassium phosphate pH 7.4; and HPL in 10 mM sodium phosphate pH 7.9. This dilution of ionic strength was necessary to decrease the absorption contribution of salt in the "far" UV regions, particularly at wavelengths below 200 nm. 30 µl of sample was inserted into a sample cell with a 0.02 mm pathlength, and secured in a sample holder. The required concentration of protein was determined using the following equation:

Equation 6.1:

$$p = \frac{0.1}{c}$$

Where p = the cell pathlength, c = concentration of protein, and 0.1 refers to the cell pathlength used during CSA calibration.

Data were collected over a complete range of 170 to 260 nm, to incorporate both the "far" and "near" UV regions, with an increment of 0.5 nm and a dwell period of 1.0 second. The scan resolution was set to 1 nm to maximise the light passing through the cell at each increment. Scans were repeated three times for Mtb-CYP125 and twice for HPL. Blank spectra containing buffer alone were performed before each sample scan.

The blank data were subtracted from the sample, before normalisation using the CSA calibration curve, by an in-house data reduction program. The web-based program, Dichroweb was then used to analyse the data (Lobley *et al.*, 2002, Whitmore and Wallace, 2004, and http://www.cryst.bbk.ac.uk/cdweb/html/home.html). Three algorithms were used during data analysis: CDSSTR (Compton and Johnson, 1986 and Sreerama and Woody, 2000); SELCON3 (Sreerema and Woody, 1993 and Sreerema *et al.*, 1999); and CONTINLL (Provencher and Glockner, 1981 and Van Stokkum *et al.*, 1990). Although all three methods have been found to be comparable (Sreerama and Woody, 2000), it was thought reliability may be improved by comparing outputs.

The resulting CD data plots for both Mtb-CYP125 and HPL are shown in **figures 6.12** and **6.13**. Only one dataset for each protein is shown graphically: Mtb-CYP125 (sample 3, **table 6.4**); and HPL (sample 2, **table 6.4**). The agreement between experimental and optimised datasets is more accurate using CDSSTR for both proteins. This program fits the experimental data against an optimised set of CD data points within the database. A comparison of the experimental data for Mtb-CYP125 and HPL is shown in **figure 6.14**.

The resulting secondary structural predictions from an average of all three algorithms (and from CDSSTR alone) for Mtb-CYP125 are: ~ 33 % ± 3.0 (31 % ± 1.2) α-helix; ~ 14 % ± 2.6 (14 % ± 2.5) β-sheet; ~ 19 % ± 1.3 (19 % ± 1.0) loops/turns; and ~ 34 % ± 2.0 (34 % ± 2.0) unordered.

The resulting secondary structural predictions from an average of all three algorithms (and from CDSSTR alone) for HPL are: ~ 45 % ± 5.0 (48 % ± 2.0) α-helix; ~ 9 % ± 6.0 (5 % ± 2.0) β-sheet; ~ 17 % ± 3.2 (18 % ± 0.5) loops/turns; and ~ 29 % ± 2.4 (29 % ± 1.0) unordered.

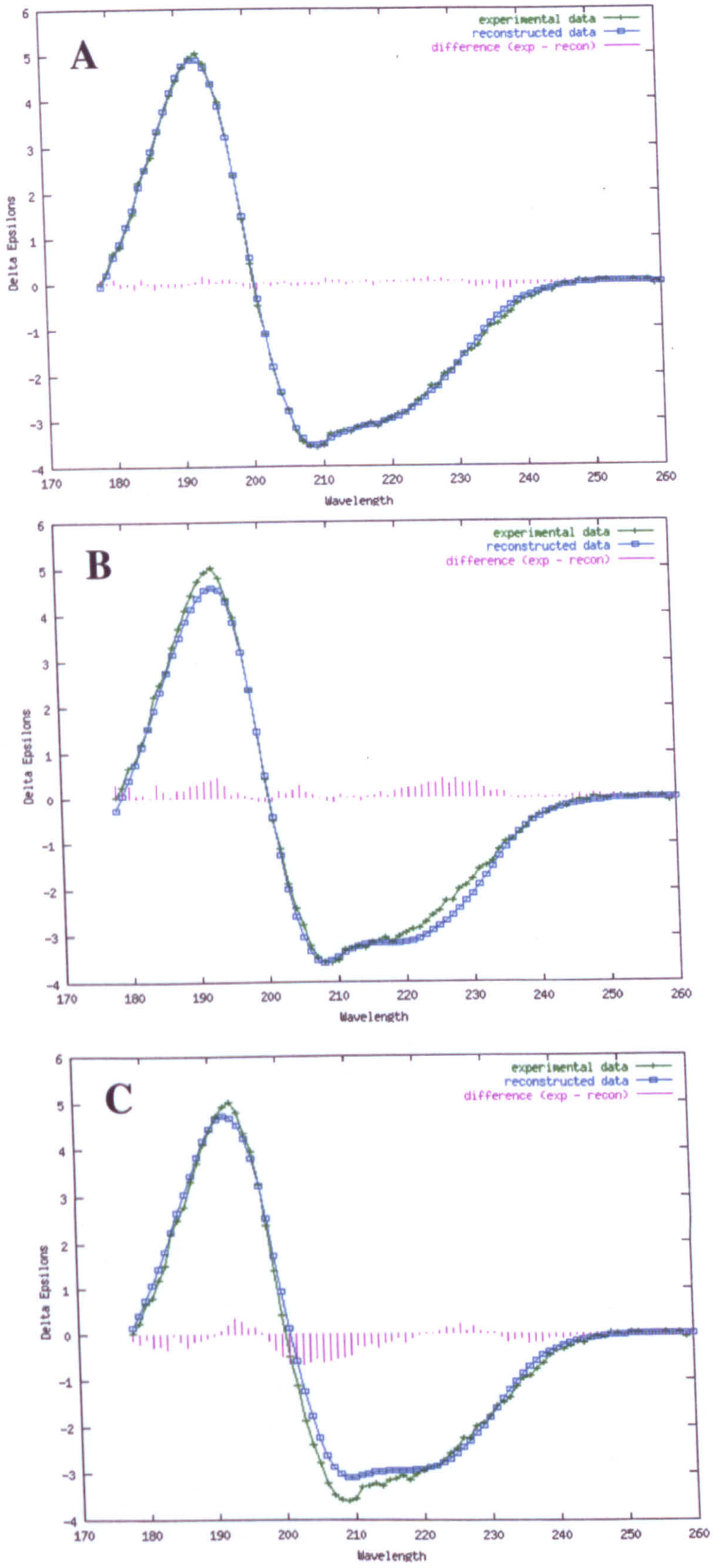


Figure 6.12: Graphical output of the experimental (green), optimised (blue), and difference (pink) spectra from CD datasets of Mtb-CYP125 (sample 3, **table 6.4**). The optimised datasets were determined by: **(A)** CDSSTR; **(B)** CONTINLL; and **(C)** SELCON3.

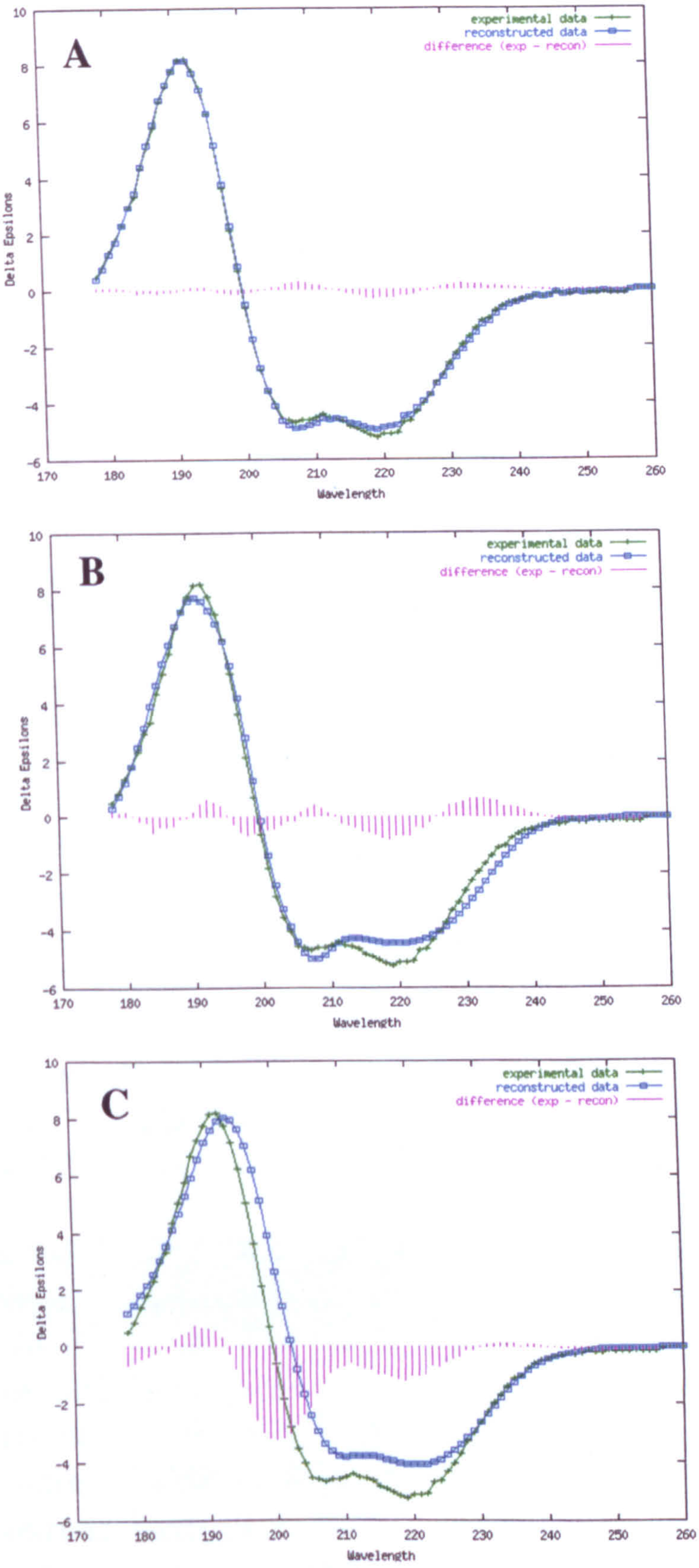


Figure 6.13: Graphical output of the experimental (green), optimised (blue), and difference (pink) spectra from CD datasets of HPL (sample 2, **table 6.4**). The optimised datasets were determined by: (**A**) CDSSTR; (**B**) CONTINLL; and (**C**) SELCON3.

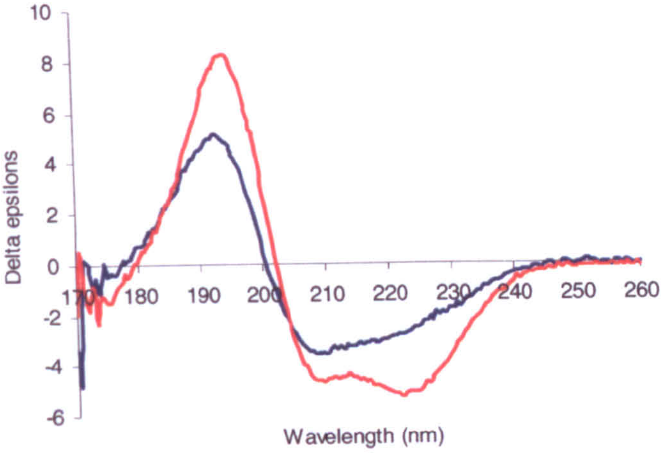


Figure 6.14: Comparison of the experimental CD spectra for Mtb-CYP125 (blue) and HPL (red).

Mtb-CYP125 sample	Analysis method	α-helix	310-helix	β-sheet	Turns	Loops	Unordered	Total
1	CDSSTR	0.220	0.080	0.114	0.110	0.090	0.360	1.000
	CONTINLL	0.255	0.091	0.120	0.121	0.082	0.330	0.999
	SELCON3	0.232	0.080	0.125	0.135	0.054	0.340	0.965
2	CDSSTR	0.248	0.068	0.154	0.132	0.055	0.337	0.993
	CONTINLL	0.250	0.070	0.150	0.130	0.070	0.340	1.001
	SELCON3	0.253	0.072	0.148	0.127	0.067	0.334	1.001
3	CDSSTR	0.250	0.070	0.150	0.120	0.070	0.330	0.990
	CONTINLL	0.268	0.076	0.137	0.127	0.068	0.325	1.001
	SELCON3	0.250	0.066	0.158	0.131	0.053	0.338	0.996
AVERAGE		**0.247**	**0.079**	**0.140**	**0.126**	**0.068**	**0.337**	**0.994**
AVERAGE (%)		**24.7**	**7.9**	**14.0**	**12.6**	**6.8**	**33.7**	**99.4**

HPL sample	Analysis method	α-helix	310-helix	β-sheet	Turns	Loops	Unordered	Total
1	CDSSTR	0.360	0.100	0.070	0.140	0.030	0.300	1.000
	CONTINLL	0.317	0.091	0.123	0.121	0.061	0.288	1.001
	SELCON3	0.355	0.088	0.098	0.116	0.045	0.314	1.016
2	CDSSTR	0.370	0.130	0.030	0.170	0.010	0.280	0.990
	CONTINLL	0.317	0.101	0.111	0.140	0.050	0.280	0.999
	SELCON3	0.386	0.075	0.118	0.102	0.036	0.269	0.986
AVERAGE		0.351	0.098	0.091	0.132	0.039	0.289	0.999
AVERAGE (%)		**35.1**	**9.8**	**9.1**	**13.2**	**3.9**	**28.9**	**99.9**

Table 6.4: Analysis from CD measurements of Mtb-CYP125 and HPL from *Medicago truncatula* using CDSSTR, CONTINLL, and SELCON3 via Dichroweb (http://www.cryst.bbk.ac.uk/cdweb).

6.5 Electron paramagnetic resonance (EPR)

6.5.1 Introduction

Electron paramagnetic resonance (EPR) was used to confirm the spin-state of Mtb-CYP125, determined by comparative analysis of UV/visible spectra with published data (section 6.3.2a). Both substrate-free and substrate-bound systems were measured in the presence and absence of imidazole, respectively.

6.5.2 Methods and results

EPR spectra were collected using a Bruker ESP300E spectrometer fitted with an X-band microwave bridge (SuperX, ER 049X), dielectric resonator (ER4118 SPT-N1), and a variable temperature liquid helium flow cryostat (Oxford Instruments auto-tuning temperature controller ITC503). Spectra were obtained at the EPSRC National Centre for EPR Spectroscopy, Manchester, with the help of Dr. Radoslaw Kowalczyk. Mtb-CYP125 (substrate-bound) was purified as described in section 4.3.2e and concentrated to 96 mg/ml (2 mM) in 50 mM potassium phosphate pH 7.4, 500 mM potassium chloride, and 1 mM dithioreitol. Data were recorded at 10 K, with a microwave power of 200 mW and a frequency of 9.43 GHz. The sample was thawed and 300 mM imidazole was added, resulting in an inhibited, substrate-free system (section 6.3.2a). Data were recorded at 30 K, with a microwave power of 50 mW, at a frequency of 9.38 GHz, due to overloading of the signal in the low-spin region at lower temperatures. g-values were estimated using the following calculation:

Equation 6.2:

$$g = (h/\mu B)\ v / B$$

Where:

h = Planck's constant

μ = Bohr's magnetron

B = the experimental magnetic field

v = frequency

The value of (h/μB) during these experiments was 714.4775.

At 10 K, the substrate-bound Mtb-CYP125 exhibited a spectrum characteristic of a predominantly high-spin (S = 5/2) haem iron system, with corresponding g-values at 8.05, 3.56, and 1.68 (**figure 6.15**). Inclusion of the reversible inhibitor, imidazole, which binds to the distal haem iron position, resulted in a spin-shift to a low-spin (S = 1/2) system, with g-values at 2.47, 2.27, and 1.88 (**figure 6.16**).

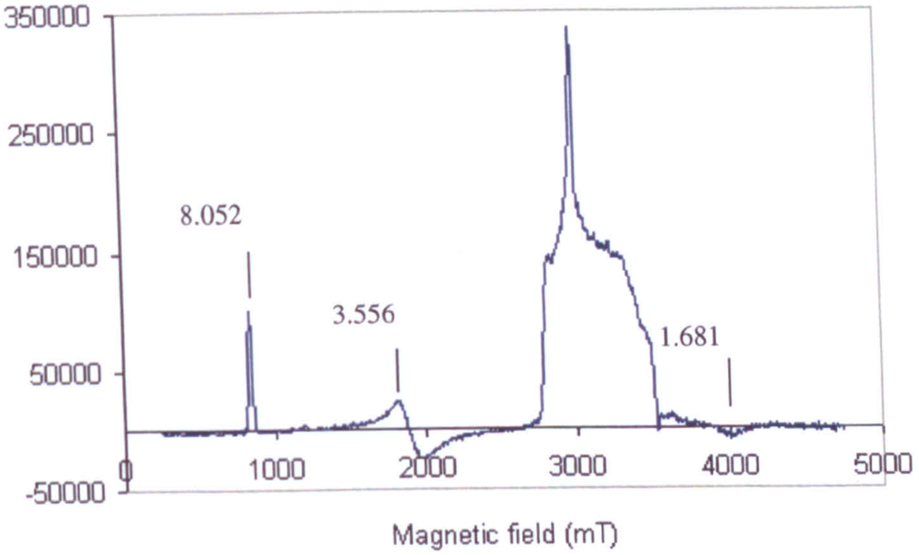


Figure 6.15: EPR spectrum of Mtb-CYP125 (substrate-bound), measured at 10 K. The corresponding g-values, characteristic of a high-spin haem iron system, are labelled.

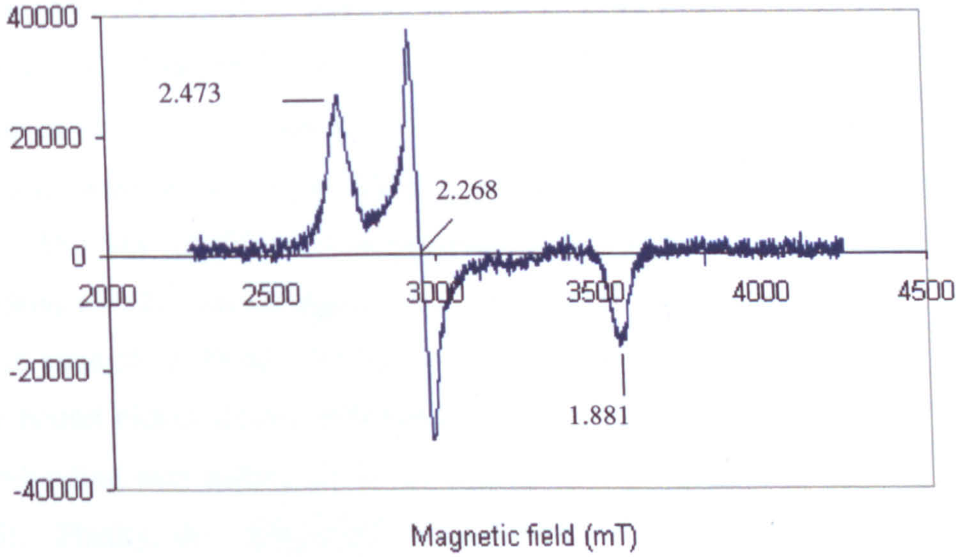


Figure 6.16: EPR spectrum of Mtb-CYP125 (substrate-free) inhibited by imidazole (300 mM), measured at 30 K. The corresponding g-values, characteristic of a low-spin haem iron system, are labelled.

6.6 Crystallisation

6.6.1 Introduction

Despite the abundance of cytochrome P450s within nature, only 169 structures are available within the Protein Data Bank, 45 of which are various forms of the most structurally characterised P450, P450cam. P450s are often difficult to crystallise, particularly the membrane-bound enzymes, which often require engineering to remove transmembrane domains in order to facilitate solubility prior to crystallisation.

Robotic crystallisation was used to identify initial hits, to enable a large number of conditions to be screened rapidly. A smaller number of manual screens were also performed, and potential hits were optimised manually.

6.6.2 Methods and results

Mtb-CYP125 was expressed and purified as described in section 4.3.2e and concentrated using Vivaspin-6 30 kDa centrifuge filtration units. Purity was determined spectroscopically with a ratio of 392 nm (Mtb-CYP125, HS) to 280 nm of > 1.0. The protein was centrifuged for 5 minutes at 14, 000 rpm using a bench-top Eppendorf centrifuge at 4 °C, immediately prior to crystallisation, to remove precipitate. All screens were set up at room temperature and incubated at 20 °C. Buffer conditions and protein concentrations were varied (**table 6.5**), to increase the number of different conditions screened. The HS gel-filtration preparation of Mtb-CYP125 was screened at three concentrations (10, 20, and 40 mg/ml) in high-salt buffer (B – D, **table 6.5**), and also at a lower ionic strength at 20 and 40 mg/ml (F – G, **table 6.5**). Crystallisation trials of the imidazole-bound nickel-affinity preparation (LS) were also performed, as it was thought imidazole-binding may induce a conformational shift conducive to crystal formation (A, **table 6.5**). Finally, the crystallisation conditions from three Mtb-CYP125 homologues were also tested, in buffers H to J (**table 6.5**).

Preparation ID	Final chromatography step	Concentration mg/ml	Buffer
A	Nickel-affinity	20	500 mM KPi pH 7.4 300 mM Imidazole
B	Gel filtration	40	50 mM KPi pH 7.4 500 mM KCl
C	Gel filtration	20	50 mM KPi pH 7.4 500 mM KCl
D	Gel filtration	10	50 mM KPi pH 7.4 500 mM KCl
E	Gel filtration	20	50 mM KPi pH 7.4 20 % Glycerol 250 mM KCl
F	Gel filtration	40	50 mM KPi pH 7.4 150 mM KCl
G	Gel filtration	20	50 mM KPi pH 7.4 150 mM KCl
H	Gel filtration	40	50 mM KPi pH
I	Gel filtration	10	50 mM Tris-HCl pH 7.4
J	Gel filtration	10	10 mM Tris-HCl pH 7.5 50 mM NaCl

Table 6.5: Mtb-CYP125 preparations and buffer conditions used for protein crystallisation.

6.6.2a Robotic screening

The broad matrix 96-well screens which were available in our laboratory, were used to identify initial hits (**table 6.6**) (Qiagen, formerly Nextal Biotechnologies). With the exception of H to J, all buffer/sample conditions described in **table 6.5** were screened against these precipitant conditions. Due to the limited quantity of protein retained from the nickel-affinity preparation (A, **table 6.5**), only two 96-well screens were performed using this sample. These were the JCSG and PEGS screens, which gave promising results using the gel-filtration Mtb-CYP125 preparations (**table 6.7** and **figure 6.17**).

Screen	Preparation ID[1]
Nextal $AmSO_4$	B/C/E/F/G
Nextal Cations	B/C/E/F/G
Nextal Classics	B/C/E/F/G
Nextal Cryo	B/C/E/F/G
Nextal JCSG	A/B/C/E/F/G
Nextal MPD	B/C/E/F/G
Nextal PACT	B/C/E/F/G
Nextal PEGS	A/B/C/E/F/G

Table 6.6: Robotic broad-matrix crystallisation conditions used for initial screening of Mtb-CYP125. [1]The purification and buffer conditions of the protein, as described in **table 6.5**.

200 nl of protein was mixed with an equal volume of precipitant over an 80 µl reservoir, in a 96-well sitting-drop plate, using a Screenmaker 96 + 8 (Innovadyne Technologies) robot. Plates were then covered with a heat-sealable plastic sheet and viewed using a Crystal Pro robot with Crystal L.I.M.S. software (both Tritek Corporation) at regular intervals.

A number of hits were obtained using the PEGS, JCSG, and PACT screens, in the form of brown/red clusters of small crystalline plates, which appeared after approximately one week (**table 6.7** and **figure 6.17**). Colourless crystals were disregarded due to the intense colour of Mtb-CYP125 in solution. A requirement for PEG was observed in all hits but the overall crystal morphology was largely unaffected by changes in salt content, and no single crystals were observed in any condition. The most promising condition from the PEG screen was that of 20 % PEG 3350 with 0.2 M ammonium chloride, which gave large clusters of plates. Slightly less compacted clusters were obtained when using the lower protein concentration of 20 mg/ml for this condition. Increasing the molecular weight of PEG to 6000 and including 0.1 M MES pH 6.0, gave better defined and less densely packed clusters of crystals, although again no single crystals were visible.

Hit number	Preparation ID[1]	Screen	Well	Condition	Description
1	E	PEGS	E5	20 % PEG 3350 0.2 M Magnesium chloride	Multiple tightly packed clusters of dark brown needles
2	C	PEGS	E9	20 % PEG 3350 0.2 M Ammonium chloride	Tightly packed cluster of dark brown needles
3	B	PEGS	E9	20 % PEG 3350 0.2 M Ammonium chloride	2 tightly packed clusters of dark brown needles
4	C	PEGS	E12	20 % PEG 3350 0.2 M Ammonium iodide	Multiple clusters of brown needles
5	B	JCSG	A5	20 % PEG 3350 0.2 M Magnesium formate	Multiple tightly packed clusters of dark brown needles
6	C	PACT	B7	0.1 M MES pH 6.0 20 % PEG 6000 0.2 M Sodium chloride	Large cluster of brown/pink needles
7	C	PACT	B8	0.1 M MES pH 6.0 20 % PEG 6000 0.2 M Ammonium chloride	Large cluster of brown/pink needles
8	C	PACT	B9	0.1 M MES pH 6.0 20 % PEG 6000 0.2 M Lithium chloride	Medium clusters of small brown/pink needles
9	C	PACT	B10	0.1 M MES pH 6.0 20 % PEG 6000 0.2 M Magnesium chloride	Less densely packed clusters of brown/pink needles

Table 6.7: Main hits obtained from the robotic screening of crystallisation conditions for Mtb-CYP125 using broad-matrix Nextal screens. [1]The purification and buffer conditions of the protein (**table 6.5**). Refer to **figure 6.17** for photographs of crystal hits.

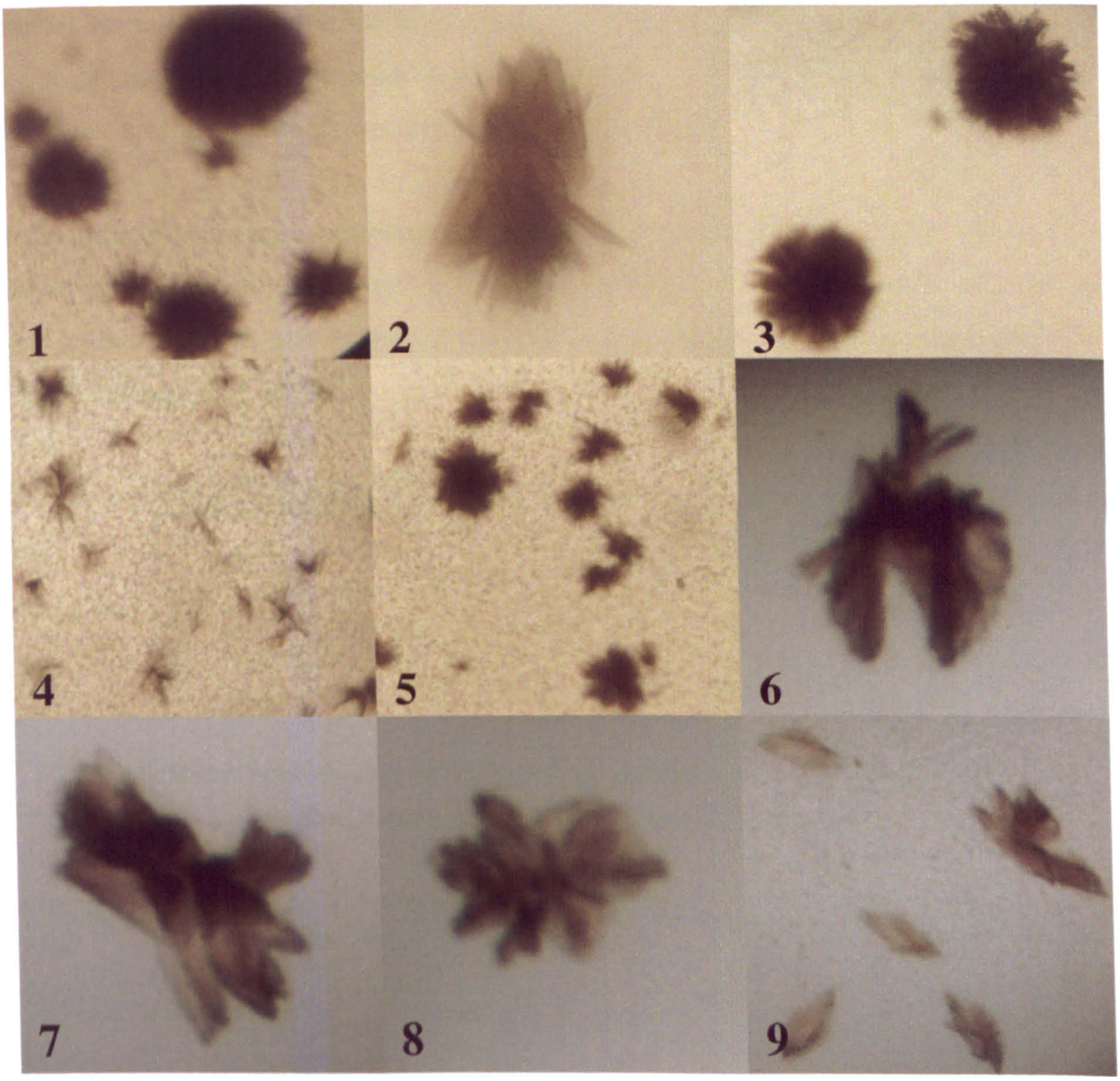


Figure 6.17: Mtb-CYP125 crystals obtained from robotic screening, see **table 6.7** for descriptions.

6.6.2b Manual screening

Manual screening of suitable conditions for Mtb-CYP125 crystallisation were performed using a standard 24-well pre-greased plate, suitable for hanging-drop vapour-diffusion crystallisation (VDX plate, Hampton Research). 1 to 2 µl of protein was mixed, using a pipette, with an equal volume of precipitant on a siliconised cover slip, suspended over a 500 µl reservoir and incubated at 20 °C. Both commercially sourced (Hampton Research) and hand-made screens based upon conditions used to crystallise several homologues were performed (**tables 6.8** and **6.9**). Commercial screens were selected if they included conditions which had not been screened during the robotic trials. Only a limited number of manual screens were performed as coloured crystals were identified in the robotic trials, and so optimisation of these conditions were prioritised. Mtb-CYP125 was screened at 10 mg/ml, as a starting concentration for the crystallisation trials.

No significant hits were produced using the screens described in **tables 6.8** and **6.9**. However small colourless crystals were produced from a number of these conditions, but were presumed to be salt as they did not show the deep red/brown colour of the Mtb-CYP125 protein.

Preparation ID[1]	Screen
D	Hampton Research Crystal Screen 1
D	Hampton Research Crystal Screen 2
D	Hampton Research Cryo
D	Hampton Research Salt RX 1
D	Hampton Research Sodium Malonate

Table 6.8: Manual broad-matrix crystallisation screens used to identify hits for Mtb-CYP125. [1]The purification and buffer conditions of the protein (**table 6.5**).

Preparation ID[1]	Screen	Protein	Species	Sequence identity with Mtb-CYP125	PDB ID
H	0.1 M Pipes pH 6.4 – 6.8 12 – 24 % PEG 12 k	P450terp	*Pseudomonas sp.*	29 %	1CPT[2]
I	0.1 M MES pH 5.0 – 6.0 0.8 - 3.2 M Ammonium sulphate	CYP121	*Mycobacterium tuberculosis*	27 %	1N40[3]
J	0.1 M Sodium cacodylate pH 6.2 – 6.5 18 – 25 % PEG 4 k 10 % Isopropanol	CYP51 α-sterol methylase	*Mycobacterium tuberculosis*	23 %	1H5Z[4]

Table 6.9: Manual crystallisation screens used during crystallisation trials of Mtb-CYP125, based upon conditions used to crystallise homologous proteins. [1]The purification and buffer conditions of Mtb-CYP125 (**table 6.5**). [2]Hasemann *et al.*, 1994. [3]Leys *et al.*, 2003. [4]Podust *et al.*, 2004.

6.6.2c Optimisation

A number of hits, obtained by robotic screening, were optimised by fine screening of the conditions. Predominantly this was done using the manual hanging-drop method, but one screen was set up in a 96-well sitting-drop plate, using a Microlab STARlet (Hamilton Research) liquid handler and a Screenmaker 96 + 8 (Innovadyne Technologies) robot.

Precipitant, salt, and pH were varied during this step, together with the concentration of protein used and the ratio of protein to precipitant (**table 6.10**). In the absence of crystals in some wells, small clusters of crystals from the robotic screens were used to seed fresh drops after 5 to 7 days. Clusters were either crushed using a needle and streak-seeded, or transferred intact to a drop using a loop.

Screening around the initial conditions did not visibly improve crystal morphology and still resulted in the growth of multiple crystals (**table 6.11** and **figure 6.18**). A slight improvement was observed however by increasing the ammonium chloride concentration to 0.4 M in the presence of 22 % PEG 3350, which yielded crystals less densely packed than previously (optimisation number 1, **table 6.11**). Increasing the molecular weight of PEG to 20 k also had a similar effect. In all screens performed, crystals only grew spontaneously when fresh protein (stored at 4 °C for < 3 weeks, following purification) was used. In some cases, seeding could induce the growth of crystals after this time, although after approximately 4 weeks, no crystallisation was observed.

Optimisation number	Hit number[1]	Preparation ID[2]	Screen	Method[3]	Ratio[4]
1	2	C	15 – 25 % PEG 3350 0.1 – 0.4M Ammonium chloride	H-D	1 : 1
2	2	C	15 – 25 % PEG 3350 0.1 – 0.4 M Ammonium chloride 0.1 M buffer pH 4.6 (NaAc), 6.5 (MES), 7.5 (Hepes), 8.4 (Tris-HCl)	H-D	1 : 1
3	2	C	15 – 24 % PEG 3350 0.1 – 0.4 M Ammonium chloride 0.1 – 0.4 M buffer pH 6.5 (MES), 7.5 (Hepes), 8.5 (Tris-HCl)	S-D	1 : 1
4	2	C	20 – 25 % PEG 3350 0.3 – 0.6 M Ammonium chloride	H-D	1 : 1
5	2	C	12 – 25 % PEG 3350 0.1 – 0.6 M Ammonium chloride	H-D	1 : 2
6	3	B	15 – 24 % PEG 3350 0.1 – 0.4 M Ammonium chloride 0.1 – 0.4 M buffer pH 6.5 (MES), 7.5 (Hepes), 8.5 (Tris-HCl)	S-D	1 : 1
7	4	C	15 – 25 % PEG 3350 0.1 – 0.4 M Ammonium iodide	H-D	1 : 1
8	4	C	15 – 25 % PEG 3350 0.1 – 0.4 M Ammonium iodide 0.1 M buffer pH 4.6 (NaAc), 6.5 (MES), 7.5 (Hepes), 8.4 (Tris-HCl)	H-D	1 : 1
9	4	B	15 – 24 % PEG 3350 0.1 – 0.4 M Ammonium chloride 0.1 – 0.4 M buffer pH 6.5 (MES), 7.5 (Hepes), 8.5 (Tris-HCl)	S-D	1 : 1
10	4	C	15 – 24 % PEG 3350 0.1 – 0.4 M Ammonium chloride 0.1 – 0.4 M buffer pH 6.5 (MES), 7.5 (Hepes), 8.5 (Tris-HCl)	S-D	1 : 1
11	7	B	19 – 24 % PEG 6k/8k/120k/20k 0.1 – 0.4 M Ammonium chloride 0.1 M MES pH 6.0	H-D	1 : 1
12	7	C	19 – 24 % PEG 6k/8k/120k/20k 0.1 – 0.4 M Ammonium chloride 0.1 M MES pH 6.0	H-D	1 : 1

Table 6.10: Optimisation of Mtb-CYP125 crystallisation conditions from initial robotic screening hits. [1]Robotic screening hit (**table 6.7**). [2]Purification and buffer conditions of the protein (**table 6.5**). [3]The crystallisation method used, with “H-D” specifying manual hanging-drop vapour diffusion and “S-D”, robotic sitting-drop vapour diffusion. [4]The ratio of protein to precipitant.

Hit number	Optimisation number[1]	Preparation ID[2]	Condition	Seeded[3]	Description
10	1	C	22 % PEG 3350 & 0.4 M Ammonium chloride	No	Less densely packed clusters of brown/red needles
11	5	C	23 % PEG 3350 & 0.6 M Ammonium chloride	Yes	Thin clusters of brown needles
12	12	C	21 % PEG 12k, 0.2 M Ammonium chloride & 0.1 M MES pH 6.0	No	Multiple clusters of dark brown needles
13	12	C	19 % PEG 20k, 0.2 M Ammonium chloride & 0.1 M MES pH 6.0	No	Less densely packed clusters of brown/pink needles
14	12	C	22 % PEG 20k, 0.2 M Ammonium chloride & 0.1 M MES pH 6.0	No	Thin clusters of brown/pink needles
15	12	C	23 % PEG 20k, 0.2 M Ammonium chloride & 0.1 M MES pH 6.0	No	Small clusters of brown/pink needles

Table 6.11: Most improved hits from optimisation of Mtb-CYP125 crystallisation conditions. All obtained by manual hanging-drop vapour diffusion crystallisation. [1]The optimisation screen (**table 6.10**). [2]The purification and buffer conditions of the protein (**table 6.5**). [3]Crystals grown after streak-seeding with crushed needle clusters after 7 days (seeds obtained from optimisation number 4, **table 6.7**). Photographs of the crystals are shown in **figure 6.18**.

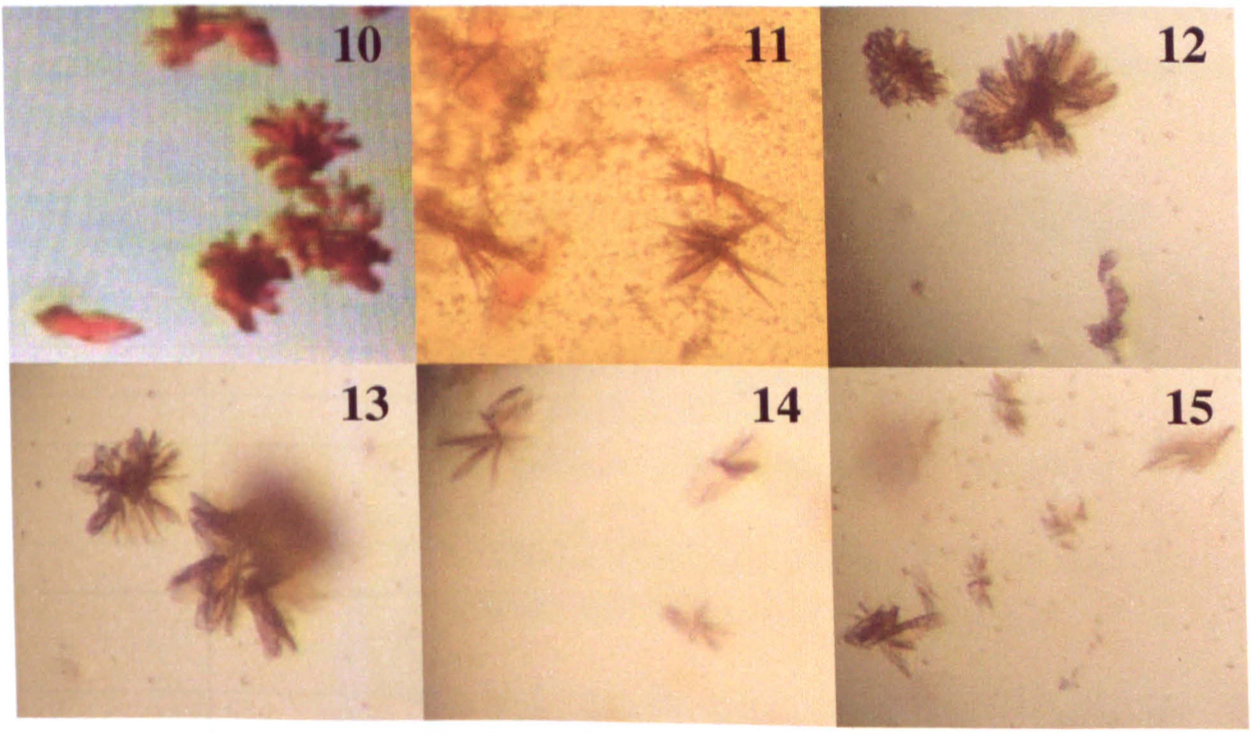


Figure 6.18: Mtb-CYP125 crystals obtained from manual optimisation of robotic hit numbers 2 and 3 (**table 6.7**). No significant change in crystal morphology was observed, although the number of plates per cluster were somewhat reduced. Descriptions of crystal conditions are given in **table 6.11**.

6.6.2d X-ray crystallography data collection

The most promising (least densely packed) clusters of crystals were used for X-ray data collection on station MAD 10.1 at the Synchrotron Radiation Source, Daresbury. These are listed in **table 6.12**. In the absence of single crystals, it was necessary to attempt to separate the multiple crystals using a needle. Crystals were briefly soaked in a cryoprotectant of mother liquor containing 20 % glycerol, mounted using a cryo-loop, and flash-cooled to 100 K in a nitrogen cryostream. An X-ray diffraction pattern was collected using an exposure time of 90 seconds, from each crystal. Data were collected over an oscillation range of 1.0 °, at a wavelength of 1.17 Å, with the crystal to detector distance set to 300 mm.

Only very weak diffraction was observed from all of the crystals, at a maximum resolution of 3 Å for hit number 2 (**table 6.12**). It was not possible to separate the crystals entirely and so multiple-crystal diffraction was also observed. In an attempt to improve diffraction, the cryostream was interrupted briefly to enable annealing of the crystal to occur. This produced a slightly more defined diffraction pattern for hit number 7 (**table 6.12**), to a maximum resolution of 3 Å (**figure 6.19**), as identified using HKL2000 (Otwinowski and Minor, 1997), however it was not possible to reproducibly determine unit cell parameters. Additional cryoprotectants were also tested, however these did not improve the diffraction quality.

Hit number[1]	Preparation ID[2]	Condition
2	C	20 % PEG 3350 & 0.2 M Ammonium chloride
3	B	20 % PEG 3350 & 0.2 M Ammonium chloride
5	B	20 % PEG 3350 & 0.2 M Magnesium formate
6	C	20 % PEG 6k, 0.2 M Sodium chloride & 0.1 M MES pH 6.0
7	C	20 % PEG 6k, 0.2 M Ammonium chloride & 0.1 M MES pH 6.0
8	C	20 % PEG 6k, 0.2 M Lithium chloride & 0.1 M MES pH 6.0
9	C	20 % PEG 6k, 0.2 M Magnesium chloride & 0.1 M MES pH 6.0
10	C	22 % PEG 3350 & 0.4 M Ammonium chloride
11	C	23 % PEG 3350 & 0.6 M Ammonium chloride

Table 6.12: Mtb-CYP125 crystals used during X-ray data collection. [1]Robotic screening hit (**tables 6.7 & 6.11** and **figures 6.17 & 6.18**). [2]Purification and buffer conditions of the protein (**table 6.5**).

Figure 6.19: The weak diffraction pattern generated from multiple Mtb-CYP125 crystals grown in 20 % PEG 6000, 0.2 M ammonium chloride, and 0.1 M MES pH 6.0 (hit number 7, **table 6.12**). The whole image is shown in **(A)** and **(B)** shows a region close to the maximum resolution limit of 3 Å, where a number of very weak diffraction spots are just visible. Figure produced using HKL2000 (Otwinowski and Minor, 1997).

6.7 Discussion

6.7.1 Comparison of P450 sequences

From the homology searches performed in section 6.2.2a, CYP125 was identified in two other non-*Mycobacterial* species, *Rhodococcus sp.* (RHA1) and *Nocardiodes sp.* (JS614), the latter shares only 55 % sequence identity with Mtb-CYP125. The unknown protein from *Mycobacterium paratuberculosis* may share the same functional annotation as Mtb-CYP125 due to the high sequence identity between the two proteins. No three-dimensional structures of proteins from the CYP125 family exist to date, highlighting the need for structural information regarding Mtb-CYP125.

Alignment of Mtb-CYP125 with P450 homologues of known structure identified conserved regions of catalytic and architectural importance, most notably around the Cys-loop. This strong conservation suggests the region within Mtb-CYP125 will exhibit a structure similar to that found in other P450s, such as P450eryF (Cupp-Vickery *et al.*, 2001). The identification of a threonine residue (Thr272) within the probable I-helix region of Mtb-CYP125 suggests this enzyme does not require a water molecule to stabilise the oxy-ferryl bond, as is the case for P450eryF (Cupp-Vickery and Poulos, 1995 and Poulos *et al*, 1995). Furthermore, conservation in Mtb-CYP125, of polar/charged residues involved in haem coordination in a number of P450s, also suggests a similar configuration (see section 1.3.7g). The lack of homology in regions implicated in substrate access and binding is common in P450s, reflecting the large number of different substrates metabolised by these enzymes.

6.7.2 Secondary structure

Circular dichroism was used to corroborate secondary structural information gained from bioinformatics methods. Prediction software such as PredictProtein can identify potential localised structure for an idealised sample. CD, however, cannot be used to identify specific regions, but does provide a more accurate representation of the protein in its existing state. This partially explains the differences observed between the two methods, whereby CD (all algorithms) and PreditProtein calculated Mtb-CYP125 to contain 33 and 41 % α-helix and 14 and 8 % β-sheet, respectively (**table 6.4**). Apart from these differences, both suggest a predominantly helical structure, which is to be expected for a cytochrome P450 (Poulos *et al.*, 1987 and Ravichandran *et al.*, 1993).

Analysis of the CD measurements using three algorithms, calculated HPL to contain an average of ~ 45 % α-helix and just ~ 9 % β-sheet. **Figure 6.14** clearly shows two troughs, characteristic of helical structure, present at ~ 210 and 225 nm within the HPL sample, which are much less prominent within Mtb-CYP125. Increased helical content has also been observed in two Mtb-CYP125 homologues: greater than 50 % in Mtb-CYP121 (McLean and Cheesman *et al.*, 2002); and ~ 44 % in CYP119 from *S. solfataricus* (Maves and Sligar, 2006), further suggesting that the CD-determined helical content of Mtb-CYP125 is lower than expected. Partial degradation of the sample during storage at 4 °C may account for some of this discrepancy, as the content of unordered structure (~ 34 %) was slightly higher than in the HPL sample (~ 29 %).

An alternative explanation is described in section 1.3.8, where Yun *et al.* (1996) identified an increased α-helical content of CYP1A2 in the presence of increasing ionic strength. Initial measurements in the absence of sodium chloride yielded just ~ 30 % α-helix content, similar to the result obtained for Mtb-CYP125, which was also measured in an low salt environment (50 mM potassium phosphate, pH 7.4). Inclusion of 0.1 M sodium chloride in the CYP1A2 sample increased the value to ~ 49 %. This may account for the low, CD-determined, helical content of Mtb-CYP125.

6.7.3 Effect of "substrate" on spin-state

The ability of Mtb-CYP125 to alternate between spin-states, in otherwise identical buffer systems, upon reversible inhibition by imidazole, is of interest due to the unknown nature of its substrate. Since the main buffer component, phosphate, was found not to have an effect on spin state when diluted out, whatever compound is mimicking a natural substrate, must bind to the protein with high affinity to remain within the system when not bound during affinity-chromatography. Alternative suggestions are that the "substrate" is retained from this purification step and binds to Mtb-CYP125 during desalting (of imidazole) by gel-filtration, or that the "substrate" binds to a co-purifying protein. In either case it seems unlikely that this compound represents a native substrate, unless a complex was formed within the *E. coli* host cell during expression. The action of phosphate on altering the spin state cannot, however, be ruled out entirely and may require competitive dilution by another compound.

Ionic strength is known to play an important role in the catalytic activity of some P450s (Yun *et al.*, 1996) and high concentrations of sodium chloride have been found to stabilise these enzymes by preferential hydration (Timasheff and Arakawa, 1989). Spectral shifts, indicating alterations in spin-state, have been observed in a number of P450s in the presence of high salt concentrations (Yun *et al.*, 1998). However, removal of both phosphate and potassium chloride from the high-spin Mtb-CYP125, did not result in a return to the resting, low-spin system. This suggests either strong "substrate" binding or that potassium exerts little or no effect on the binding of substrate to Mtb-CYP125. As an effect of potassium on substrate binding affinity has, to date, only been found in P450cam, the latter seems most likely (Peterson, 1971, Poulos *et al.*, 1987, and Mueller *et al.*, 1995).

EPR measurements of Mtb-CYP125 (substrate-bound) resulted in a spectrum characteristic of a predominantly high-spin haem iron system. The g-values (8.05 g_z, 3.56 g_y, 1.68 g_x) were consistent with those published for P450cam in the presence of D-camphor (7.85, 3.97, 1.78) (Tsai *et al.*, 1970 and Lipscomb, 1980). Similar g-values were also obtained from the HPL spectrum, however this sample was deemed to be substrate-free by the authors (8.03, 3.51, 1.68) (Hughes *et al.*, 2006).

Inclusion of imidazole shifted the spin state to low-spin, with g-values (2.47, 2.27, 1.88), similar to the substrate-free low-spin P450cam (2.45, 2.26, 1.91), P450BM-3 (2.42, 2.26, 1.96) (Miles *et al.*, 1992), and Mtb-CYP121 (2.48, 2.25, 1.90) (McLean *et al.*, 2005). Again these values correspond to those identified in the low-spin HPL system, however substrate was present in this sample (2.39, 2.24, 1.93) (Hughes *et al.*, 2006). Overall, these data confirm the Mtb-CYP125 spin-states of substrate-bound (high-spin) and substrate-free (low-spin) systems, as determined by UV/visible spectroscopy.

6.7.4 CO-binding assay

Together with the UV/visible spectra of ferric Mtb-CYP125 described in section 6.3.2a, the CO-binding assay of ferrous protein confirmed the cytochrome P450 annotation of this enzyme. In both the substrate-free and substrate-bound forms, a distinct peak at 450 nm was observed upon anaerobic addition of CO. The shift was especially prominent in the substrate-bound system, due to the shift in Soret peak from 392 nm to 450 nm. In both cases, reduction with dithionite did not alter the maximal absorbance wavelength significantly, however a large decrease in intensity was observed in the substrate-bound protein, resulting in a much broader peak, as is characteristic of P450s (McLean and Cheesman *et al.*, 2002).

The presence of a second peak at 420 nm for both forms of Mtb-CYP125 was initially thought to mean partial degradation of the protein during purification. Glycerol is known to confer limited protection against such conversion and so was included during the purification of additional protein. A repeat CO-binding assay found less P420 conversion than in the glycerol-free preparation, however did not eliminate it entirely.

Further CO-binding experiments performed using Mtb-CYP125, in the absence of glycerol, showed a time-dependent increase in the production of both P420- and P450-CO complexes. P420 often forms in the ferrous state, possibly due to protonation of the cysteinate –S$^-$ group, which forms a neutral cysteine, and so is unlikely to have formed within the oxidised protein (Perera *et al.*, 2002). This suggests the issue was a limitation of the experiment and not an inherent problem with the protein itself.

Also, full conversion to P420 was observed upon loss of the proximal cysteinate ligand in the P450cam C357H mutant (section 1.3.7a), which was not apparent in the Mtb-CYP125 spectra (Yoshioka *et al.*, 2001). This suggests that the partial conversion observed for Mtb-CYP125 is of less significance. Furthermore, characterisation of the Mtb-CYP121 enzyme observed aggregation and precipitation upon addition of large excesses of dithionite, and partial conversion to P420 was noted in the presence of CO (McLean and Cheesman *et al.*, 2002). The subsequent crystal structure of ferric Mtb-CYP121 found an intact cysteinate ligand, demonstrating the cause of P420 conversion to be entirely due to the CO-binding experiment (Leys *et al.*, 2003).

6.7.5 Protein crystallography

Despite obtaining dark brown/red Mtb-CYP125 crystals from a number of crystallisation conditions, it was not possible to collect a full data set. This was primarily due to the morphology of the crystals, tiny plates clustered together into groups of varying size. This made separation of individual crystals impossible and despite extensive optimisation of the conditions, no alternate crystal morphology has been observed. The large conformational changes which may occur during substrate binding may affect crystal morphology, and so crystallisation of a substrate-free molecule would be of interest. The inclusion of such a high concentration of imidazole, needed to displace the substrate, is likely to interfere with crystallisation (Li and Poulos, 2004), however a number of P450 structures have been successfully determined in the presence of azole compounds (Yano *et al.*, 2000, Scott *et al.*, 2004, and Verras *et al.*, 2006).

Although the weak diffraction from multiple crystals, described in section 6.6.2d, could not unambiguously establish unit cell parameters, the presence of diffraction up to 3 Å is promising. Since the functional annotation of Mtb-CYP125 has been confirmed through spectroscopic analysis, a structure would provide significant information, particularly in the presence of azole inhibitors.

Chapter 7 – Overall conclusions and future work

The aim of the research presented in this thesis was to overproduce a number of uncharacterised metalloproteins from *Mycobacterium tuberculosis*, and to provide structural information regarding these targets. Initially, a cell-free system was exploited to express multiple targets, primarily for its high-throughput capabilities, and was performed at the RIKEN Yokohama Institute due to their continued success in this area (Kigawa, 1999 and 2002, Yokoyama, 2003, and Matsuda *et al.*, 2006). This method was used to express milligram quantities of soluble protein for 9 out of the 28 targets, in six weeks. It is unlikely that comparable results would have been obtainable using an *in vivo* system in the short timeframe, unless a high-throughput system was in place (Murthy *et al.*, 2004). Limited optimisation of the reaction conditions, for targets which did not express or were insoluble, were also performed, however no improvements were identified.

A second 12 week visit to RIKEN focused on the expression of 8 new targets, together with a more thorough optimisation of expression conditions for several insoluble targets. Soluble expression was obtained for four of the new targets, two predicted zinc-binding proteins, one iron-binding protein, and one unknown metalloprotein. Optimisation of the remaining targets proved highly successful, with partial solubility obtained for all proteins upon addition of detergents or molecular chaperones to the cell-free reactions. However the yield of soluble protein in these cases was often low and problems arose during removal of these additives. It seems that where possible, it is beneficial to express protein without the inclusion of such additives.

Expression of four targets using the *in vivo E. coli* method gave similar results to those obtained by the cell-free method, however solubility of Mtb-CYP125 (Rv3545c) was significantly improved. Negligible levels of soluble protein were obtained for this target using *in vivo* expression conditions similar to those used to express homologous cytochrome P450s, whereby soluble protein was obtained 6 to 24 hours following induction (Bellamine *et al.*, 1999 and McLean and Cheesman *et al.*, 2002). Increased incubation times of 72 hours were required to produce significant quantities of soluble Mtb-CYP125, however the reason for this remains unknown.

The success rate of synthesising soluble haem proteins using the cell-free system is reported to be low, due to problems associated with the incorporation of the large prosthetic group (Matsuda, personal communication). Due to the presence of whole cells within the *in vivo* system, it was possible to include a smaller precursor within the expression media, which was then converted to the functional haem group.

The cell-free system described in this thesis provides a rapid method for the high-throughput identification of target solubility. For structural genomics projects, this is clearly advantageous as soluble targets can quickly be identified and progressed into large-scale synthesis, in preparation for downstream applications. A turnover of less than two days can be achieved from target isolation to analysis of protein expression using this system, whilst a typical (non-high-throughput) *E. coli in vivo* method would require approximately double this. However, negative aspects associated with the cell-free system include the potential high cost to set up and the requirement for high-grade reagents and cell extracts (Murthy *et al.*, 2004). Commercially sourced plasmids and host cells, together with standard molecular biology laboratory equipment, can be used to set up an *in vivo* system with little complication.

Overall, the cell-free system was extremely successful at producing soluble protein for the metalloproteins targeted. However similar results were obtained, albeit on a smaller scale, using the *in vivo* system. In the absence of a dedicated *in vitro* protein expression facility and a constant supply of high-grade cell-free components, the *in vivo* system appears to be preferential but time-consuming. It would be useful to express a larger number of targets using the *in vivo* system, in order to provide a more detailed comparison. Of particular interest would be a detailed comparison of the timescales required, number of soluble targets produced and purification steps required, yield following purification, and biological activity.

The 2.7 Å resolution structure of Mtb-PPase (Rv3628) showed an overall monomeric fold characteristic of prokaryotic type I PPases. Despite intra-trimer interactions involving poorly-conserved residues, the oligomeric form also remained conserved. Substrate/product and metal binding is well documented in Y-PPase (Harutyunyan *et al.*, 1996 and Heikinheimo *et al.*, 1996) and E-PPase (Harutyunyan *et al.*, 1997 and Samygina *et al.*, 2001), with little variation exhibited between the two enzymes. Although no metal ions were identified within the Mtb-PPase structure, one phosphate group was modelled at site P1, forming bonds with Arg37, Tyr133, and Lys134. Homologous interactions exist in both the yeast and *E. coli* structures. The conservation of all 17 active site residues (Sivula *et al.*, 1999) and the similarity in overall structure and P1 binding site of Mtb-PPase with homologous structures, suggests metal coordination and the P2 binding site will also be similar.

Superimposition of key active site residues of Mtb-PPase with the calcium-inhibited E-PPase structure (Samygina *et al.*, 2001) showed a marked difference in the position and orientation of Glu25, Asp59, and Asp96. Since the calcium ions in E-PPase locate the activating metal sites, such differences are expected as these residues are known to participate in metal coordination. Also, Lys23 was found to adopt a more extended conformation in the E-PPase structure, forming a hydrogen bond with the P2 site. This phosphate group was not present in the Mtb-PPase structure and so explains this structural difference.

Comparison with two recent Mtb-PPase structures, both in space group $P6_322$ (Tammenkoski *et al.*, 2005 and Benini and Wilson, to be published), highlighted a possible pH-dependent role of His93 within the active site. Mutation of this residue was found to only hamper activity in the presence of magnesium (Tammenkoski *et al.*, 2006) and so whether His93 is of catalytic importance, in the presence of other activating metal ions, remains unclear.

Evidence for a critical role of PPase in both *E. coli* (Chen *et al.*, 1990) and *S. cerevisiae* (Lundin *et al.*, 1991) suggest Mtb-PPase may be an attractive target for therapeutic intervention and future development of specific inhibitors. Although three structures of Mtb-PPase now exist, none represent the catalytically active enzyme, and so crystallisation in the presence of activating metals and phosphate/pyrophosphate, would be of interest. It also remains of high priority to determine the H-PPase structure in order to identify structural differences between the two enzymes, so that therapeutic Mtb-PPase inhibitors can be realised. Potential inhibitors may target non-active site residues which have essential catalytic roles in Mtb-PPase, but are not required in H-PPase.

Characterisation of the Rv3545c gene product (Mtb-CYP125) confirmed its functional annotation as a cytochrome P450. This target was chosen for expression and characterisation studies because of its essential nature during the *in vivo* infection of mice with *M. tb*, together with our group's ongoing interest in cytochromes. Sequence homology searches identified only two further non-*Mycobacterial* CYP125s in the UNIPROT database and no crystal structures currently exist for this protein. Alignment of Mtb-CYP125 with P450 homologues of known structure identified a number of conserved residues within the Cys-loop, including the explicitly conserved cysteine residue, which provides the proximal thiolate ligand to the haem iron. Also of interest is the identification of a threonine residue (Thr272) within the probable I-helix, which suggests Mtb-CYP125 does not require a water molecule to stabilise the oxy-ferryl bond, as is the case for P450eryF (Cupp-Vickery and Poulos, 1995 and Poulos *et al*, 1995).

Bioinformatics and circular dichroism (CD) were used to calculate the secondary structural elements of Mtb-CYP125. CD measurements showed the enzyme to contain 33 % α-helix and 14 % β-sheet. The α-helical content deviates somewhat from the expected value for P450s of > 40 % (McLean and Cheesman *et al.*, 2002 and Maves and Sligar, 2006). The low ionic strength at which Mtb-CYP125 CD data were collected may account for this (Yun *et al.*, 1996) and so further experiments conducted at varying salt concentration would be useful. The prediction software, PredictProtein, estimated Mtb-CYP125 to contain 41 % α-helical structure, further suggesting that the CD-derived value may be inaccurate.

Routine spectroscopic analysis of the protein following affinity chromatography and gel filtration identified an interesting Soret shift between the two purification steps. Further investigation found this shift to be caused by the reversible dilution of the inhibitor, imidazole, during gel filtration. The resulting spectrum after gel filtration was characteristic of P450 in a high-spin system, with substrate blocking the access of water to the distal haem iron ligand (Li, 2001), and was confirmed by subsequent EPR measurements. A physiological substrate of Mtb-CYP125 remains unknown and due to the vast number of substrates metabolised by the P450 superfamily, no compound has yet been inferred from sequence homology. Unless a native substrate was encountered in the cell during expression and remained in the system throughout purification, it remains unclear why this shift is observed. There is clearly a necessity to determine the crystal structure of this enzyme to determine what, if any, compound is bound to the active site. A successful method for the expression and purification of Mtb-CYP125 is documented in this thesis and further optimisation of the crystallisation conditions may produce crystals suitable for high resolution X-ray diffraction. Also of interest, given the potential of Mtb-CYP125 to act as a novel drug target, would be a crystal structure complexed with inhibitors such as azole compounds.

Tuberculosis is a devastating disease and much remains unknown about its pathogenicity and how it evades the host immune system with such efficiency. Given the current reliance on five chemotherapeutics and the resultant widespread resistance, it is crucial that new drug targets are identified and characterised. The high-throughput cell-free expression system described in this thesis proved highly successful for the rapid identification of soluble targets, which could then be synthesised on a large-scale to produce milligram quantities of protein. This may represent a viable option for the production of multiple targets in preparation for downstream applications, enabling the characterisation of potential drug targets, and ultimately leading to the production of novel antimycobacterials.

References

Ahn, S., Milner, A. J., Futterer, K., Konopka, M., Ilias, M., Young, T. W., and White, S. A. (2001) The "open" and "closed" structures of the type-C inorganic pyrophosphatases from *Bacillus subtilis* and *Streptococcus gordonii*. *Journal of Molecular Biology*. **313**: 797-811.

Aikens, J. and Sligar, S. G. (1994) Kinetic solvent isotope effects during oxygen activation by cytochrome P-450cam. *Journal of the American Chemical Society*. **116**: 1143-1144.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. Molecular Biology of the Cell. 316-317. Garland Publishing, New York, USA (1994).

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*. **215**:403-410.

Anderson, D. H., Harth, G., Horwitz, M. A., and Eisenberg, D. (2001) An interfacial mechanism and a class of inhibitors inferred from two crystal structures of the Mycobacterium tuberculosis 30 kDa major secretory protein (Antigen 85B), a mycolyl transferase. *Journal of Molecular Biology*. **307**: 671-681.

Andrews, A. T. "Electrophoresis of Nucleic Acids" in Essential Molecular Biology: A Practical Approach. Volume 1. 89 – 126. Edited by Brown, T. A. IRL Press, Oxford, UK. 1992.

Arutiunian, E. G., Terzian, S. S., Voronova, A. A., Kuranova, I. P., Smirnova, E. A., Vainstein, B. K., Hohne, W. E., and Hansen, G. (1981) X-Ray diffraction study of inorganic pyrophosphatase from baker's yeast at the 3 Angstroms resolution (article in Russian). *Dokl.Akad.Nauk SSSR*. **258**: 148.

Avaeva, S. M, Kurilova, S., Nazarova, T., Rodina, E., Vorobyeva, N., Sklyankina, V., Grigorjeva, O., Harutyunyan, E., Oganessyan, V., Wilson, K. (1997) Crystal structure of *E. coli* inorganic pyrophosphatse complex with SO_4^{2-}. *FEBS Letters*. **410**: 502-508.

Avaeva, S. M., Rodina, E. V., Vorobyeva, N. N., Kurilova, S. A., Nazarova, T. I., Sklyankina, V. A., Oganessyan, V. Y., and Harutyunyan, E. H. (1998) Changes in *E. coli* inorganic pyrophosphatase structure induced by binding of metal activators. *Biochemistry (Moscow).* **63**: 592-599.

Baranov, V. I., Morozov, I. Y., Ortlepp, S. A., and Spirin, A. S. (1989) Gene expression in a cell-free system on the preparative scale. *Gene.* **84**: 463-466.

Baranov, V. I., Ryabova, L. A., Yarchuk, O. B., and Spirin, A. S. (2002) Method of preparing polypeptides in a cell-free translation system. PTC Filed, June 14, 1990. *United States. Patent #* 6,399,323 B1.

Baykov, A. A. and Shestakov, A. S. (1992) Two pathways of pyrophosphate hydrolysis and synthesis by yeast inorganic pyrophosphatase. *European Journal of Biochemistry.* **206**: 463-470.

Beale, D., and Feinstein, A. (1976) Structure and function of the constant regions of immunoglobulins. *Quarterly Reviews of Biophysics.* **9**: 135-180.

Bellamine, A., Mangla, A. T., Nes, W. D., and Waterman, M. R. (1999) Characterization and catalytic properties of the sterol 14α-demethylase from Mycobacterium tuberculosis. *Proceedings of the National Academy of Science USA.* **96**(16): 8937-8942.

Benini, S. and Wilson, K.S. The crystal structure of *Mycobacterium Tuberculosis* inorganic pyrophosphatase. *Recent Advances in Inorganic Pyrophosphatase Research,* presented at the Proceedings of the Third International Meetings on Inorganic Pyrophosphatases. 25-31. Edited by White, S. A. Birmingham, UK. 2004.

Benini, S. and Wilson, K.S. *Mycobacterium Tuberculosis* Rv3628, Yet Another Inorganic Pyrophosphatase or a Possible Drug Target. To be Published.

Berman, H. M., Henrick, K., Nakamura, H. (2003) Announcing the worldwide Protein Data Bank. *Nature Structural Biology.* **10**(12): 980.

Blundell, T. L. and Johnson, L. N. Protein Crystallography. Academic Press, London, UK. 1976.

Boshoff, H. I. M., Myers, T. G., Copp, B. R., McNeil, M. R., Wilson, M. A., and Barry, C. E. (2004) The transcriptional responses of *Mycobacterium tuberculosis* to inhibitors of metabolism: NOVEL INSIGHTS INTO DRUG MECHANISMS OF ACTION. *Journal of Biological Chemistry.* **279**: 40174-40184.

Boshoff, H. I. M. and Barry, C. E. (2005) Tuberculosis - metabolism and respiration in the absence of growth. *Nature Reviews Microbiology.* **3**: 70-80.

Bossi, R.T., Aliverti, A., Raimondi, D., Fischer, F., Zanetti, G., Ferrari, D., Tahallah, N., Maier, C. S., Heck, A. J. R., Rizzi, M., Mattevi, A. (2002) A covalent modification of NADP+ revealed by the atomic resoution structure of FprA, a *Mycobacterium tuberculosis* oxidoreductase. *Biochemistry.* **41**: 8807-8818.

Bradford, M. (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry.* **72**: 248-254.

Bragg, W. L. (1912) The diffraction of short electromagnetic waves by a crystal. *Proceedings of the Cambridge Philosophical Society.* **17**: 43–57.

Brennan, P. J. (1995) Biogenesis of the mycobacterial cell wall and the site of action of ethambutol. *Antimicrobial Agents and Chemotherapy.* **11**: 2484-89.

Brinkmann, U., Mattes, R. E., and Buckel, P. (1989) High level expression of recombinant genes in *Escherichia coli* is dependant on the availability of the *dnaY* gene product. *Gene.* **85**(1): 109–114.

Brown, T. A. "The Essential Techniques in Molecular Biology" in Essential Molecular Biology: A Practical Approach. Volume 1. 1 – 11. Edited by Brown, T. A. IRL Press, Oxford, UK. 1992.

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

Busam, R. D., Thorsell, A., Flores, A., Hammarström, M., Persson, C, and Hallberg, B. M. (2006) First Structure of a eukaryotic phosphohistidine phosphatase. *Journal of Biological Chemistry*. **281**: 33830-33834.

Butler, L. G. Yeast and other inorganic pyrophosphatases in The Enzymes (3rd edition). Volume 4. Edited by Boyer, P. D. Academic Press, New York, USA. 1971.

Butler, L. G. and Sperow, J. W. (1977) Multiple roles of metal ions in the reaction catalyzed by yeast inorganic pyrophosphatase. *Bioinorganic Chemistry*. **7**: 141-150.

Carter, A. P., Clemons, W. M., Brodersen, D. E., Morgan-Warren, R. J., Wimberly, B. T., and Ramakrishnan, V. (2000) Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature*. **407**: 340-348.

Collaborative Computational Project, Number 4. (1994) The CCP4 suite: Programs for protein crystallography. *Acta Crystallographica D*. **50**: 760-763.

Chapple, C. (1998) Molecular-genetic analysis of plant cytochrome P450-dependant monooxygenases. *Annual Review of Plant Molecular Biology*. **49**: 311-343.

Chen, J., Brevet, A., Fromant, M., Lévêque, F., Schmitter, J. M., Blanquet, S., and Plateau, P. (1990) Pyrophosphatase is essential for growth of *Escherichia coli*. *Journal of Bacteriology*. **172**: 5686-5689.

Chumpolkulwong, N., Sakamoto, K., Hayashi, A., Iraha, F., Shinya, N., Matsuda, N., Kiga, D., Urushibata, A., Shirouzu, M., Oki, K., Kigawa, T., and Yokoyama, S. (2006) Translation of 'rare' codons in a cell-free protein synthesis system from *Escherichia coli*. *Journal of Structural and Functional Genomics*. **7**: 31-36

Cianci, M., Antonyuk, S., Bliss, N., Bailey, M., Buffey, S., Cheung, K., Clarke, J., Derbyshire, G., Ellis, M., Enderby, M., Grant, A., Holbourn, M., Laundy, D., Nave, C., Ryder, R., Stephenson, P., Helliwell, J., Hasnain, S. (2005) A high-throughput structural biology/proteomics beamline at the SRS on a new multipole wiggler. *Journal of Synchrotron Radiation.* **12**: 455-466.

Cohen, S. N., Chang, A. C. Y., and Hsu, L. (1972) Non-chromosomal antibiotic resistance in bacteria: Genetic transformation of *Escherichia coli* by R-factor DNA. *Proceedings of the National Academy of Science.* **69**: 2110-2114.

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., Tekaia, F., Badcock, K., Bxam, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L, Oliver, K., Osborne, J., Quail, M., A., Rajandream, M-A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J. E., Taylor, K., Whitehead, S., Barrell, B. G. (1998) Deciphering the Biology of Mycobacterium tuberculosis from the Complete Genome Sequence. *Nature.* **393**: 537-544.

Cole, S. T., Camus, J-C., Pryor, M. J., Medigue, C. (2002) Re-Annotation of the Genome Sequence of Mycobacterium tuberculosis H37Rv. *Microbiology.* **148**: 2967-2973.

Collman, J. P. and Sorrell, T. N. (1975) A model for the carbonyl adduct of ferrous cytochrome P-450. *Journal of the American Chemical Society.* **97**: 4133-4134.

Compton, L. A. and Johnson, W. C., Jr. (1986) Analysis of protein circular dichroism spectra for secondary structure using a simple matrix multiplication. *Analytical Biochemistry.* **155**: 155-167.

Cooperman, B. S., Panackal, A., Springs, B., and Hamm, D. J. (1981) Divalent metal ion, inorganic phosphate, and inorganic phosphate analogue binding to yeast inorganic pyrophosphatase. *Biochemistry.* **20**: 6051-6060.

Cooperman, B. S., Baykov, A. A., and Lahti, R. (1992) Evolutionary conservation of the active site of soluble inorganic pyrophosphatase. *Trends in Biochemical Sciences*. **17**(7): 262-266.

Correia, M. A. and Ortiz de Montellano, P. R. "Inhibition of Cytochrome P450 Enzymes" in Cytochrome P450 Structure, Mechanism, and Biochemistry (3rd edition). Edited by Ortiz de Montellano, P. R. Plenum Press, New York, USA. 2005.

Crowther, R. A. and Blow, D. M (1967) A method of positioning a known molecule in an unknown crystal structure. *Acta Crystallographica*. **23**: 544-548.

Cruickshank, D. W. J. Protein precision re-examined: Luzzati plots do not estimate final errors. Proceedings of CCP4 Study Weekend: Refinement of Macromolecular Structures (Chester). 1996.

Cupp-Vickery, J. R., Li, H., and Poulos, T. L. (1994) Preliminary crystallographic analysis of an enzyme involved in erythromycin biosynthesis: Cytochrome P450eryF. *Proteins*. **20**(2): 197-201.

Cupp-Vickery, J. R. and Poulos, T. L. (1995) Structure of cytochrome P450eryF involved in erythromycin biosynthesis. *Nature Structural Biology*. **2**: 144-153.

Cupp-Vickery, J. R., Garcia, C., Hofacre, A., and McGee-Estrada, K. (2001) Ketaconazole-induced conformational changes in the active site of Cytochrome P450eryF. *Journal of Molecular Biology*. **311**: 101-110.

Dalvi, R. R. (1987) Cytochrome P-450 dependant covalent binding of carbon dioxide to rat liver microsomal protein *in vitro* and its prevention by reduced glutathionie. *Archives of Toxicology*. **61**: 155-157.

Dauter, Z., Dauter, M., De La Fortelle, E., Bricogne, G., and Sheldrick, G. M. (1999) Can anomalous signal of sulfur become a tool for solving protein crystal structures? *Journal of Molecular Biology*. **289**: 83-92.

Davies, H. S., Britt, S. G., and Pohl, L. R. (1986) Carbon tetrachloride and 2-isopropyl-4-pentenamide-induced inactivation of cytochrome P-450 leads to heme-derived protein adducts. *Archives in Biochemistry and Biophysics.* **244**: 352-387.

Davis, B. J. (1964) Disc electrophoresis II: Method and application to human serum proteins. *Annals of the New York Academy of Science.* **121**: 404-427.

Del Tito, B. J. (1995) Effects of a minor isoleucyl tRNA on heterologous protein translation in *Escherichia coli. Journal of Bacteriology.* **177**: 7086–7091.

Drenth, J. Principles of Protein X-ray Crystallography (2nd edition). Springer, New York, USA. 1999.

Deprez, E., Di Primo, C., Hoa, G. H., and Douzou, P. (1994). Effects of monovalent cations on cytochrome P-450 camphor: Evidence for preferential binding of potassium. *FEBS Letters.* **347**: 207-210.

Dessen, A., Guilmi, A. D., Vernet, T., and Dideberg, O. (2005) Molecular mechanisms of antibiotic resistance in gram-positive pathogens. Bentham Sciences Publishers. Online publication.

Dickins, M., Elcombe, C. R., Moloney, S. J., Netter, K. J., and Bridges, J. W. (1979) Further studies on the dissociation of the isoafrole metabolite-cytochrome P-450 complex. *Biochemistry and Pharmacology.* **28**: 231-238.

Dombradi, V. (2002) Structure and function of protein phosphatases. *European Journal of Biochemistry.* **269**: 1049–1049.

Dressen, A., Guilmi, A. M., Vernet, T., and Dideberg, O. (2005) Molecular mechanisms of antibiotic resistance in gram-positive pathogens. http://www.bentham.org.

Ducruix, D. and Giege, R. Crystallisation of Nucleic Acids and Proteins: A Practical Approach. Oxford University Press, Oxford, UK. 1992.

Dunn, M. J. "Initial Planning" in Protein Purification Methods: A Practical Approach. 10 - 39. Edited by Harris, E. L. V. and Angal, S. IRL Press, Oxford, UK. 1989.

Durst, F. and Nelson, D. R. (1995) Diversity and evolution of plant P450s and P450-reductases. *Drug Metabolism and Drug Interactions.* **12**: 189-206.

Efimova, I. S., Salminen, A., Pohjanjoki, P., Lapinniemi, J., Magretova, N. N., Cooperman, B. S., Goldman, A., Lahti, R., and Baykov, A. A. (1999) Directed mutagenesis studies of the metal binding site at the subunit interface of *Escherichia coli* inorganic pyrophosphatase. *Journal of Biological Chemistry.* **274**: 3294-3299.

Ellis, M. J., Antonyuk, S. A., and Hasnain, S. S. (2002) Resolution improvements from '*in situ* annealing' of copper nitrite crystals. *Acta Crystallographica D.* **58**: 456-458.

Emsley, P. and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallographica D.* **60**: 2126-2132.

Engh, R. A. and Huber, R. (1991) Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallographica A.* **47**: 392-400.

Evans, J. (1998) TB: Know Your Enemy. *Chembytes E-Zine.* The Royal Society of Chemistry. http://www.chemsoc.org/chembytes/ezine/1998/evans.htm

Fabrichniy, I. P., Lehtio, L., Salminen, A., Zyryanov, A. B., Baykov, A. A., Lahti, R., and Goldman, A. (2004) Structural studies of metal ions in family II pyrophosphatases: The requirement for a janus ion. *Biochemistry.* **43**: 14403-14411.

Falzon, M., Nielsch, A., and Burke, M. D. (1986) Denaturation of cytochrome P-450 by indomethacin and other non-steroidal anti-inflammatory drugs: Evidence for a surfactant mechanism and a selective effect of a p-chlorophenyl moiety. *Biochemical Pharmacology.* **35**: 4019-4024.

Feldmann, K. A. (2001) Cytochrome P450s as genes for crop improvement. *Current Opinion in Plant Biology.* **4**: 162-167.

Fontana, E., Dansette, P. M., and Poli, S. M. (2005) Cytochrome P450 enzymes mechanism based inhibitors: Common sub-structures and reactivity. *Current Drug Metabolism.* **6**(5): 413-454.

Fox, G. C. and Holmes, K. C. (1968) An alternate method for solving the layer scaling equation of Hamilton, Rollett and Sparks. *Acta Crystallographica.* **20**: 886-891.

Fraichard, A., Trossat, C., Perotti, E., and Pugin, A. (1996) Allosteric regulation by Mg2+ of the vacuolar H(+)-PPase from Acer pseudoplatanus cells. Ca2+/Mg2+ interactions. *Biochimie.* **78**: 259-266.

French, G. S. and Wilson, K. S. (1978) On the treatment of negative intensity observations. *Acta Crystallographica A.* **34**: 517-525.

Gan, L-S. L., Abo, A. L., and Alworth, W. L. (1984) 1-Ethynylpyrene, a suicide inhibitor of cytochrome P-450 dependant benzo(a)pyrene hydroxylase activity in liver microsomes. *Biochemistry.* **23**: 3827-3836.

Gassel, M. (1999) The KdpF subunit is part of the K(+)-translocating Kdp. Gateway Technology Manual. http://invitrogen.com/content/sfs/manuals/gatewayman.pdf

Gerber, N. C. and Sligar, S. G. (1994) A role for Asp-251 in cytochrome P-450cam oxygen activation. *Journal of Biological Chemistry.* **269**: 4260-4266.

Gewirth, D. (2003) The HKL Manual.
http://www.hkl-xray.com/hkl_web1/hkl/manual_online.pdf

Giacovazzo, C., Monaco, H. L., Artioli, G., Viterbo, D., Ferraris, G., Gilli, G., Zanotti, G., and Catti M. Fundamentals of Crystallography (2^{nd} edition). IUCr/Oxford University Press, Oxford, UK. 2002.

Gill, H. S., Pfluegl, G. M. U., Eisenberg, D. (1999) Preliminary crystallographic studies on glutamine synthfatase from *Mycobacterium tuberculosis. Acta Crystallographica D.* **55**: 865-8.

Gillespie, S. H. (2002) Evolution of drug resistance in *Mycobacterium tuberculosis*: Clinical and molecular perspective. *Antiomicrobial agents and chemotherapy*. **46**: 267-274.

Gonzalez, A. and Nave, C. (1994) Radiation damage in protein crystals at low temperature. *Acta Crystallographica D*. **50**: 874-877.

Goodwin, G. H. "Clarification and extraction" in Protein Purification Methods: A Practical Approach. 96- 99. Edited by Harris, E. L. V. and Angal, S. IRL Press, Oxford, UK. 1989.

Goulding, C. W., Parseghian, A., Sawaya, M. R., Cascio, D., Apostol, M. I., Gennaro, M. L., Eisenberg, D. (2002) Crystal structure of a major secreted protein of *Mycobacterium tuberculosis* – MPT63 at 1.5 Å Resolution. *Protein Science*. **11**: 2887-2893.

Graham, J. E. (1999) Identification of *Mycobacterium tuberculosis* RNAs synthesized in response to phagocytosis by human macrophages by selective capture of transcribed sequences SCOTS. *Proceedings of the National Academy of Sciences*. **96**: 11554-11559.

Graham, S. E. and Peterson, J. A. (1999) How similar are P450s and what can their differences teach us? *Archives of Biochemistry and Biophysics*. **369**: 24–29.

Green, D. W., Ingram, V. M., and Perutz, M. F. (1954) The structure of haemoglobin IV. Sign determination by the isomorphous replacement method. *Proceedings of the Royal Society A*. **225**: 287-307.

Grossman, T. H., Kawasaki, E. S., Punreddy, S. R., and Osburne, M. S. (1998) Spontaneous cAMP-dependent derepression of gene expression in stationary phase plays a role in recombinant expression instability. *Gene*. **209**: 95–103.

Groves, J. T. and McCluskey, G. A. (1976) Aliphatic hydroxylation via oxygen rebound. Oxygen transfer catalyzed by iron. *Journal of the American Chemical Society*. **98**: 859-891.

Guengerich, F. P. "Human Cytochrome P450 Enzymes" in Cytochrome P450 Structure, Mechanism, and Biochemistry (2nd edition). Edited by Ortiz de Montellano, P. R. Plenum Press, New York, USA. 1995.

Halpert, J. and Neal, R. A. (1980) Inactivation of purified rat liver cytochrome P-450 by chloramphenicol. *Molecular Pharmacology*. **17**: 427-434.

Hames, B. D. and Hooper, N. M. Instant Notes Biochemistry (2nd edition). 93-94. BIOS Science Publishers Ltd, Oxford, UK. 2000.

Hammes, G. G. Spectroscopy for Biological Sciences. John Wiley and Sons Inc., New Jersey, USA. 2005.

Hanson, L. K., Eaton, W. A., Sligar, S. G., Gunsalus, I. C., Gouterman, M., and Connell, C. R. (1976) Origin of the anomalous Soret spectra of carboxycytochrome P450. *Journal of the American Chemical Society*. **98**: 2672-2674.

Harris, D. and Lowe, G. (1993) Determinants of the spin state of the resting state of cytochrome P450cam. *Journal of the American Chemical Society*. **115**: 8775-8779.

Harker, D. (1956) The determination of the phases of the structure factors of non-centrosymmetric crystals by the method of double isomorphous replacement. *Acta Crystallographica*. **9**: 1-9.

Harth, G. (1999) An inhibitor of exported *Mycobacterium tuberculosis* glutamine synthetase selectivity blocks the growth of pathogenic mycobacteria in axenic culture and in human monocytes: Extracellular proteins as potential novel drug targets. *Journal of Experimental Medicine*. **189**: 1425-1436.

Hartl, F. W. and Hayer-Hartl, M. (2002) Molecular chaperones in the cytosol: form nascent chain to folded protein. *Science*. **295**: 1852-1858.

Harutyunyan, E. H., Kuranova, I. P., Vainshtein, B. K., Hohne, W. E., Lamzin, V. S., Dauter, Z., Teplyakov, A. V., and Wilson, K. S. (1996). X-ray structure of yeast inorganic pyrophosphatase complexed with manganese and phosphate. *European Journal of Biochemistry.* **23**: 220-228.

Harutyunyan, E. H., Oganessyan, V. Y., Oganessyan, N. N., Avaeva, S. M., Nazarova, T. I., Vorobyeva, N. N., Kurilova, S. A., Huber, R., and Mather, T. (1997) Crystal structure of holo inorganic pyrophosphatase from *Escherichia coli* at 1.9 Å resolution. Mechanism of hydrolysis. *Biochemistry.* **36**: *7754*-7760.

Hasemann, C. A., Ravichandran, K. G., Peterson, J. A., and Deisenhofer, J. (1994) Crystal structure and refinement of cytochrome P450terp at 2.3 A resolution. *Journal of Molecular Biology.* **236**: 1169-1185.

Hasemann, C. A., Ravichandran, K .G., Boddupalli, S. S., Peterson, J. A., and Deisenhofer, J. (1995) Structure and function of cytochrome P450: A comparative analysis of the three-dimensional structures of P450terp, P450cam, and the hemoprotein domain of P450BM3. *Structure.* **3**: 41-62.

Heikinheimo, P., Lehtonen, J., Baykov. A., Lahti, R., Cooperman, B. S., and Goldman, A. (1996). The structural basis for pyrophosphatase catalysis. *Structure.* **4**: 1491-1508.

Helling, R. B., Goodman, H. M., and Boyer, H. W. (1974) Analysis of endonuclease R·EcoRI fragments of DNA from lambdoid bacteriophages and other viruses by agarose gel electrophoresis. *Journal of Virology.* **14**(5):1235-1244.

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

Helvig, C., Alayrac, C., Mioskowski, C., Koop, D., Poullain, D., Durst, F., and Salau, J-P. (1997) Suicide inactivation of cytochrome P450 by midchain and terminal acetylenes: A mechanistic study of inactivation of a plant lauric acid v-hydroxylase. *Journal of Biological Chemistry.* **272**: 414-421.

Henderson, R. (1990) Cryo-protection of protein crystals against radiation damage in electron and X-ray diffraction. *Proceedings: Biological Sciences.* **241**: 6-8.

Hirokawa, T., Boon-Chieng, S., Mitaku, S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics.* **14**: 378-379.

Hjelmeland, L.M. (1990) Removal of detergents from membrane proteins. *Methods in Enzymology.* **182:** 277-282.

Honer, Z. (1999) Charcterization of activity and expression of isocitrate lyase in *Mycobacterium avium* and *Mycobacterium tuberculosis. Journal of Bacteriology.* **181**: 7161-7.

Hooft, R. W. W., Vriend, G., Sander, C., Abola, E. E. (1996). Errors in protein structures. *Nature.* **381**: 272-272.

Horwitz, M. A. (1995) Protective immunity against tuberculosis induced by vaccination with major extracellular proteins of *Mycobacterium tuberculosis. Proceedings of the National Academy of Sciences.* **92**: 1530-4.

Horwitz, M. A. (2000) Recombinant bacillus Calmette-Guérin (BCG) vaccines expressing the *Mycobacterium tuberculosis* 30-kDa major secretory protein induce greater protective immunity against tuberculosis than conventional BCG vaccines in a highly susceptible animal model. *Proceedings of the National Academy of Sciences.* **97**(25): 13853-13858.

Hughes, R. K., Belfield, E. J., Muthusamay, M., Khan, A., Rowe, A., Harding, S. E., Fairhurst, S. A., Bornemann, S., Ashton, R., Thorneey, R. N. F., and Casey, R. (2006) Characterization of *Medicago truncatula* (barrel medic) hydroperoxide lyase (CYP74C3), a water-soluble detergent-free cytochrome P450 monomer whose biological activity is defined by monomer–micelle association. *Biochemical Journal.* **395**: 641–652.

Hutchinson, E. G. and Thornton, J. M. (1996) PROMOTIF - A program to id entify structural motifs in proteins. *Protein Science.* **5**: 212-220.

Hyytia, T., Halonen, P., Salminen, A., Goldman, A., Lahti, R., and Cooperman, B. S. (2001). Ligand binding sites in *Escherichia coli* inorganic pyrophosphatase: effects of active site mutations. *Biochemistry*. **40**: 4645-4653.

Ichiba, T., Shibasaki, T., Iizuka, E., Hachimori, A., and Samejima, T. (1998) Cation-induced thermostability of yeast and *Escherichia coli* pyrophosphatases. *Biochemistry and Cell Biology*. **66**: 25–31.

Ishihara, G., Goto, M., Saeki, M., Ito, K., Hori, T., Kigawa, T., Shirouzu, M., Yokoyama, S. (2005) Expression of G protein coupled receptors in a cell-free translational system using detergents and thioredoxin-fused vectors. *Protein Expression and Purification*. **41**(1): 27-37.

ITQB/UNL (2006) Protein Biochemistry Folding and Stability Group. http://www.itqb.unl.pt/~gomes/research2.htm

Janson, C. A., Degani, C., and Boyer, P. D. (1979) The formation of enzyme-bound and medium pyrophosphate and the molecular basis of the oxygen exchange reaction of yeast inorganic pyrophosphatase. *Journal of Biological Chemistry*. **254**: 3743-3749.

Jefcoate, C. R. (1978) Measurement of substrate and inhibitor binding to microsomal cytochrome P-450 by optical difference spectroscopy. *Methods in Enzymology*. **52**: 258-279.

Jones, T. A., Zou, J.-Y., Cowan, S. W. and Kjeldgaard, M. (1991) Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallographica A*. **47**: 110-119.

Jones, G. R. and Clarke, D. T. (2004) Applications of extended ultra-violet circular dichroism spectroscopy in biology and medicine. *Faraday Discussions*. **126**: 223-236.

Josse, J. (1966) Constitutive inorganic pyrophosphatase of *Escherichia coli*. II. Nature and binding of active substrate and the role of magnesium. Journal of Biological Chemistry. **241**: 1948-1955.

Kankare, J., Salminen, T., Lahti, R., Cooperman, B. S., Baykov, A. A., and Goldman, A. (1996). Structure of *Escherichia coli* inorganic pyrophosphatase at 2.2 angstrom resolution. *Acta Crystallographica D.* **52**: 551-563.

Kasner, R. J. (1973) A theoretical model for the effects of local non-polar heme environments on the redox potentials in cytochromes. *Journal of the American Chemical Society.* **95**: 2674-2677.

Katzen, F., Chang, G., and Kudlicki, W. (2005) The past, present and future of cell-free protein synthesis. *Trends in Biotechnology.* **23**: 150-156.

Kigawa, T., Yabuki, T., Yoshida, Y., Tsutsui, M., Ito, Y., Shibata, T., Yokoyama, S. (1999) Cell-free production and stable-isotope labeling of milligram-quantities of proteins. *FEBS Letters.* **442**(1): 15-19.

Kigawa, T., Yamaguchi-Nunokawa, E., Kodama, K., Matsuda, T., Yabuki, T., Matsuda, N., Ishitani, R., Nureki, O., Yokoyama S. (2002) Selenomethionine incorporation into a protein by cell-free synthesis. *Journal of Structural and Functional Genomics.* **2**(1): 29-35.

Kigawa, T., Yabuki, T., Matsuda, N., Matsuda, T., Nakajima, R., Tanaka, A., Yokoyama, S. (2004) Preparation of *Escherichia coli* cell extract for highly productive cell-free protein expression. *Journal of Structural and Functional Genomics.* **5**(1-2): 63-68.

Kitada, M., Chiba, K., Kamataki, T., and Kitagawa, H. (1977) Inhibition by cyanide of drug oxidations in rat liver microsomes. *Japanese Journal of Pharmacology.* **27**: 601-608.

Klemme, J-H. and Gest, H. (1971) Regulatory properties of an inorganic pyrophosphatase from the photosynthetic bacterium *Rhodospirillum rubrum. PNAS.* **68**(4): 721-725.

Knight, W. B., Dunaway-Mariano, D., Ransom, S. C., and Villafranca, J. J. (1984). Investigations of the metal ion-binding sites of yeast inorganic pyrophosphatase. *Journal of Biological Chemistry.* **259**(5): 2886-2895.

Komberg, A. "On the metabolic significance of phosphorylytic and pyrophosphorylytic reactions" in Horizons in Biochemistry. 251-264. Edited by Kasha, H. and Pullman, P. Academic Press, New York, NY. 1962.

Kwaik, Y. A. (1998) Induced expression of the *Legionella pneumophila* gene encoding 20-kilodalton protein during intracellular infection. *Infectious Immunology.* **66**: 202-212.

Kwak, A. K. and Harb, O. S. (1999) Phenotypic modulation by intracellular bacterial pathogens. *Electrophoresis.* **20**: 2248-2258.

Ladd, M. F. C and Palmer, R. A. Structure Determination by X-ray Crystallography. Plenum Press, New York, USA. 1994.

Lange, R., Pierre, J., and Debey, P. (1980) Visible and ultraviolet spectral transitions of camphor-bound cytochrome P-450. A comprehensive study. *European Journal of Biochemistry.* **107**: 441–445.

Lamzin, V. S. W. (1993) Automated refinement of crystal structures. *Acta Crystallographica D.* **49**: 129-147.

Laemmli, U. K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriphage T4. *Nature.* **227**: 680-685.

Lahti, R. (1983) Microbial inorganic pyrophosphatases. *Microbiology Reviews.* **47**: 169-179.

Lahti, R., Kolakowski, L. F., Heinonen, J., Vihinen, M., Pohjanoksa, K., and Cooperman, B. S. (1990) Conservation of functional residues between yeast and *E. coli* inorganic pyrophosphatase. *Biochimica et Biophysica Acta.* **1038**: 338-345.

Lahti, R., Pohjanoksa, K., Pitkaranta, T., Heikinheimo, P., Salminen, T., Meyer, P., and Heinonen, J. (1990). A site-directed mutagenesis study on *Escherichia coli* inorganic pyrophosphatase. Glutamic acid-98 and lysine-104 are important for structural integrity, whereas aspartic acids-97 and -102 are essential for catalytic activity. *Biochemistry.* **29**: 5761-5766.

Lamborg, H., and Zamecnik, P. C. (1960) Amino acid incorporation into protein by extracts of *E. coli. Biochimica et Biophysica Acta.* **42**: 206-211.

Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993) PROCHECK: A program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography.* **26:** 283-291.

Leahy, D. J., Hendrickson, W. A., Aukhil, I., and Erickson, H. P. (1992) Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. *Science.* **258**: 987-991.

Lederman, M., and Zubay, G. (1967) DNA-directed peptide synthesis. I. A comparison of T2 and *Escherichia coli* DNA directed peptide synthesis in two cell-free systems. *Biochimica et Biophysica Acta.* **149**: 253-258.

Leppanen, V. M., Nummelin, H., Hansen, T., Lahti, R., Schafer, G., and Goldman, A. (1999). *Sulfolobus acidocaldarius* inorganic pyrophosphatase: structure, thermostability, and effect of metal ion in an archael pyrophosphatase. *Protein Science.* **8**: 1218-1231.

Leslie, A. G. W. (1992) Recent changes to the MOSFLM package for processing film and image plate data. CCP4 and ESF-EACMB Newsletter on Protein Crystallography. 26.

Leys, D., Mowat, C. G., McLean, K. J., Richmond, A., Chapman, S. K., Walkinshaw, M. D., and Munro, A. W. (2003) Atomic structure of *Mycobacterium tuberculosis* CYP121 to 1.06 A reveals novel features of cytochrome P450. *Journal of Biological Chemistry.* **278**: 5141-5147.

Li, H. Cytochrome P450 in Handbook of Metalloproteins 1. 267-282. Edited by Messerschmidt, A., Huber, R., Poulos, T., and Wieghardt, K. John Wiley and Sons Inc., New York, USA. 2001.

Li, H. and Poulos, T. L. (2004) Crystallization of cytochromes P450 and substrate-enzyme interactions. *Current Topics in Medicinal Chemistry.* **4**(16): 1789-802.

Lipscomb, J.D. (1980) Electron paramagnetic resonance detectable states of cytochrome P-450cam. *Biochemistry.* **9**(15): 3590-3599.

Littlefield, J. W., Keller, E. B., Gross, J., and Zamecnik, P. C. (1955) Studies on cytoplasmic ribonucleoprotein particles from the liver of the rat. *Journal of Biological Chemistry.* **217**: 111-123.

Liu, B., Bartlam, M., Gao, R., Zhou, W., Pang, H., Liu, Y., Feng, Y., and Rao, Z. (2004). Crystal structure of the hyperthermophilic inorganic pyrophosphatase from the archaeon *Pyrococcus horikoshii. Biophysical Journal.* **86**: 420-427.

Lundin, M., Baltscheffsky, H., and Ronne, H. (1991). Yeast PPA2 gene encodes a mitochondrial inorganic pyrophosphatase that is essential for mitochondrial function. *Journal of Biological Chemistry.* **266**: 12168-12172.

Marcus, A., Efron, D., and Weeks, D. P. (1974) The wheat embryo cell-free system. *Methods in Enzymology.* **30**: 749-754.

Marston, F.A.O. and Hartley, D.L. (1990) Solubilization of protein aggregates. *Methods in Enzymology.* **182:** 264-276.

Matsuda, T., Kigawa, T., Koshiba, S., Inoue, M., Aoki, M., Yamasaki, K., Seki, M, Shinozaki, K., and Yokoyama, S. (2006) Cell-free synthesis of zinc-binding proteins. *Journal of Structural and Functional Genomics.* E-publication ahead of print.

Matthews, C. K., Van Holde, K. E., and Ahern, K. G. Biochemistry. Benjamin/Cummings, San Francisco, USA. 2000.

McKinney, J. D., Honer, K., Bentrup, Z., Munoz-Elias, E. J., Miczak, A., Chen, B., Chan, W., Swenson, D., Sacchettini, J. C., Jacobs, W. R., Russell, D. G. (2000) Persistance of *Mycobacterium tuberculosis* in macrophages and mice requires the glyoxylate shunt enzyme isocitrate lyase. *Nature*. **406**: 735-8.

McMurry, T. J. and Groves, J. T. "Metalloporphyrin Models for Cytochrome P-450" in Cytochrome-P450: Structure, Mechanism, and Biochemistry. Edited by Ortiz de Montellano, P. R. Plenum Press, New York, USA. 1986.

McLean, K. J., Cheesman, M. R., Rivers, S. L., Richmond, A., Leys, D., Chapman, S. K., Reid, G. A., Price, N. C., Kelly, S. M., Clarkson, J., Smith, W. E., and Munro, A. W. (2002) Expression, purification and spectroscopic characterization of the cytochrome P450 CYP121 from *Mycobacterium tuberculosis. Journal of Inorganic Biochemistry.* **91**: 527-541.

McLean, K. J., Marshall, K. R., Richmond, A., Hunter, I. S., Fowler, K., Kieser, T., Gurcha, S. S., Besra, G. S., and Munro, A. W. (2002) Azole antifungals are potent inhibitors of cytochrome P450 mono-oxygenases and bacterial growth in mycobacteria and streptomycetes. *Microbiology*. **148**: 2937–2949.

McLean, K. J., Sabri, M., Marshall, K. R., Lawson, R. J., Lewis, D. G., Clift, D., Balding, P. R., Dunford, A. J., Warman, A. J., McVey, J. P., Quinn, A. M., Sutcliffe, M. J., Scrutton, N. S., and Munro, A. W. (2005) Biodiversity of cytochrome P450 redox systems. *Biochemical Society Transactions*. **33**: 796-801.

McPherson, A. Crystallisation of Biological Macromolecules. Cold Spring Harbour Laboratory Press, New York, USA. 1999.

McRee, D. E. and David, P R. Practical Protein Crystallography (2nd edition). Academic Press, San Diego, USA. 1999.

Meharenna, Y. T., Li, H., Hawkes, D. B., Pearson, A. G., De Voss, J., and Poulos, T. L. (2004) Crystal structure of P450cin in a complex with its substrate, 1,8-cineole, a close structural homologue to D-camphor, the substrate for P450cam. *Biochemistry*. **43**: 9487-9494.

Merckel, M. C., Fabrichniy, I. P., Salminen, A., Kalkkinen, N., Baykov, A. A., Lahti, R., and Goldman, A. (2001) Crystal structure of *Streptococcus mutans* pyrophosphatase: A new fold for an old mechanism. *Structure*. **9**: 289-297.

Miles, J. S., Munro, A. W., Rospendowski, B. N., Smith, W. E., McKnight, J., Thomson, A. J. (1992) Domains of the catalytically self-sufficient cytochrome P-450 BM-3. Genetic construction, overexpression, purification and spectroscopic characterization. *Biochemistry Journal*. **288**: 503-509.

Moe, O. A. and Butler, L. G. (1972) Yeast inorganic pyrophosphatase. II. Kinetics of Mg 2+ activation. *Journal of Biological Chemistry*. **247**: 7308-7314.

Morant, M., Bak, S., Moller, B. L., and Werck-Reichhart, D. (2003) Plant cytochromes P450: Tools for pharmacology, plant protection, and phytoremediation. *Current Opinion in Biotechnology*. **14**: 151-162.

Morikawa, T., Mizutani, M., Aoki, N., Watanabe, B., Saga, H., Saito, S., Oikawa, A., Suzuki, H., Sakurai, N., Shibata, D., Wadano, A., Sakata, K., and Ohta, D. (2006) Cytochrome P450 CYP710A encodes the sterol C-22 desaturase in *Arabidopsis* and tomato. *Plant Cell*. **18**: 1008-1022.

Mowat, C. G., Leys, D., McLean, K. J., Rivers, S. L., Richmond, A., Munro, A. W., Lombardia, M. O., Alzari, P. M., Reid, G. A., Chapman, S. K., Walkinshaw, M. D. (2002) Crystallization and preliminary crystallographic analysis of a novel cytochrome P450 from *Mycobacterium tuberculosis*. *Acta Crystallographica D*. **58**: 704-705.

Mueller, E., Loida, P. J., and Sligar, S. G. "Twenty-five Years of P450cam Research" in Cytochrome P450 Structure, Mechanism, and Biochemistry (2nd edition). Edited by Ortiz de Montellano, P. R. Plenum Press, New York, USA. 1995.

Munro, A. W., McLean, K. J., Marshall, K. R., Warman, A. J., Lewis, G., Roitel, O., Sutcliffe, M. J., Kemp, C. A., Modi, S., Scrutton, N. S., and Leys, D. (2003) Cytochromes P450: Novel drug targets in the war against multidrug-resistant Mycobacterium tuberculosis. *Biochemical Society Transactions.* **31**(3): 625-630.

Murshudov, G. N., Vagin, A. A., and Dodson, E. J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallographica D.* **53**: 240-255.

Nagano, S., Cupp-Vickery, J. R., and Poulos, T. L. (2005) Crystal structures of the ferrous dioxygen complex of wild-type cytochrome P450Eryf and its mutants, A245S and A245T: Investigation of the proton transfer system in P450Eryf. *Journal of Biological Chemistry.* **280**: 22102-22107.

Nave, C. (1995) Radiation damage in protein crystallography. *Radiation physics and chemistry.* **45**: 483-490.

Nebbia, C., Ceppa, L., Dacasto, M., and Carletti, M. (1999) Triphenyltin acetate-mediated in vitro inactivation of rat liver cytochrome P-450. *Journal of Toxicology and Environmental Health.* **25**(6): 433-447.

Nebert, D. W. and Gonzalez, F. J. (1987) P450 Genes: Structure, evolution, and regulation. *Annual Reviews of Biochemistry.* **56**: 945-993.

Nebert, D. W. and Nelson, D. R. (1991) P450 gene nomenclature based on evolution. *Methods in Enzymology.* **206**: 3-11.

Nebert, D. W., Nelson, D. R., Coon, M. J., Estabrook, R. W., Feyereisen, R., Fujii-Kuriyama, Y., Gonzalez, F. J., Guengerich, F. P., Gunsalus, I. C., and Johnson, E. F. (1991) P450 superfamily: Update on new sequences, gene mapping, and recommended nomenclature. *DNA and Cell Biology.* **10**(1): 1-14.

Nelson, D. R., Kamataki, T., Waxman, D. J., Guengerich, F. P., Estabrook, R. W., Feyereisen, R., Gonzalez, F. J., Coon, M. J., Gunsalus, I. C., Gotoh, O., and Nebert, D. W. (1993) The P450 superfamily: Update on new sequences, gene mapping, accession numbers, early trivial names of enzymes, and nomenclature. *DNA and Cell Biology.* **12**(1): 1-51.

Nelson, D. R., Koymans, L., Kamataki, T., Stegeman, J. R., Feyereisen, R., Waxman, D. J., Waterman, M. R., Gotoh, O., Coon, M. J., Estabrook, R. W., Gunsalus, I. C., and Nebert, D. W. (1996) P450 superfamily: Update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics.* **6**: 1-42.

Nelson, D. R. "Cytochrome P450 Protocols" in Methods in Molecular Biology. Volume 107. Edited by Phillips, I. R. and Shephard E. A. Humana Press, Totowa, USA. 1998.

Nielsen, K. A. and Moller, B. L. "Cytochrome P450s in Plants" in Cytochrome P450 Structure, Function, and Mechanism (3rd edition). Edited by Ortiz de Montellano, P. R. Plenum Press, New York, USA. 2005.

Nirenberg, M. W. and Matthaei, J. H. (1961) The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polynucleotides. *Proceedings of the National Academy of Science.* **47**: 1588-1602.

Novagen (2004) Competent Cells Manual.
http://www.merckbiosciences.co.uk/docs/docs/PROT/TB009.pdf

Novagen (2006) pET System Manual.
www.emdbiosciences.com/docs/docs/PROT/TB055.pdf

Novy, R. (2001) Innovations, Novagen. 12:1-3.
www.emdbiosciences.com

Oganessyana, V. Y., Kurilovab, S. A, Vorobyevac, N. N., Nazarovab, T. I., Popova, A. N., Lebedeva, A. A., Avaeva, S. M., and Harutyunyan, E. H. (1994) X-Ray crystallographic studies of recombinant inorganic pyrophosphatase from *Escherichia coli*. *Febs Letters*. **348**: 301-304.

Omura, T. and Sato, R. (1964) The carbon monoxide-binding pigment of liver microsomes. II. Solubilisation, purification, and properties. *Journal of Biological Chemistry*. **239**: 2379-2385.

Ornstein, L. (1964) Disc electrophoresis I: Background and theory. *Annals of the New York Academy of Science*. **121**: 321-349.

Ortiz de Montellano, P. R. and Correia, M. A. "Inhibition of Cytochrome P450 Enzymes" in Cytochrome P450 Structure, Mechanism, and Biochemistry (2nd edition). Edited by Ortiz de Montellano, P. R. Plenum Press, New York, USA. 1995.

Osawa, Y. and Pohl, L. R. (1989) Covalent bonding of the prosthetic heme to protein: A potential mechanism for the suicide inactivation or activation of hemoproteins. *Chemical Results in Toxicology*. **2**: 131-141.

Ost, T. W. B., Miles, C. S., Munro, A. W., Murdoch, J., Reid, G. A., and Chapman, S. K. (2001) Phenylalanine 393 Exerts Thermodynamic Control over the Heme of Flavocytochrome P450 BM3. *Biochemistry*. **40**: 13421-13429.

Otwinowski, Z. (1993) Oscillation data reduction program. Paper presented at the Proceedings of the CCP4 Study Weekend "Data Collection and Processing". 29-30.

Otwinowski, Z. and Minor, W. Processing of X-ray diffraction data collected in oscillation model in Methods in Enzymology: Macromolecular Crystallography. Carter, A. C. W. and Sweet, R. M. Academic Press, New York, USA. 1997.

Otwinowski, Z. and Minor, W. "Processing of X-ray diffraction data collected in oscillation mode" in Methods in Enzymology: Macromolecular Crystallography. **276**(A): 307-326. Carter, C. W. Jr. and Sweet, R. M., Academic Press. 1997.

Park, S. Y., Yamane, K., Adachi, S., Shiro, Y., Weiss, K. E., and Sligar, S. G. (2000) Crystallization and preliminary X-ray diffraction analysis of a cytochrome P450 (CYP119) from *Sulfolobus solfataricus*. *Acta Crystallogrica D*. **56**: 1173-1175.

Pearson, W. R. and Lipman, D. J. (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*. **85**: 2444-2448.

Pelham, H. R. B., and Jackson, R. J. (1976) An efficient mRNA-dependent translation system from reticulocyte lysates. *European Journal of Biochemistry*. **67**: 247-256.

Perera, R., Sono, M., Sigman, J. A., Pfister, T. D., Yu Lu., and Dawson, J. H. (2002) Neutral thiol as a proximal ligand to ferrous heme iron: Implications for heme proteins that lose cysteine thiolate ligation on reduction. *Proceedings of the National Academy of Sciences*. **100**(7): 3641-3646.

Peterson, J. A. (1971) Camphor binding by *Pseudomonas putida*. *Archives of Biochemistry and Biophysics*. **144**(2): 678-693.

Peterson, J. A. and Graham-Lorence, S. E. "Bacterial P450s" in Cytochrome P450 Structure, Mechanism, and Biochemistry. Edited by Oritz de Montellano, P R. Plenum Press, New York, USA. 1995.

Pinkse, M. W., Merkx, M., and Averill, B, A. (1999) Fluoride inhibition of bovine spleen purple acid phosphatase: Characterization of a ternary enzyme-phosphate-fluoride complex as a model for the active enzyme-substrate-hydroxide complex. *Biochemistry*. **38**: 9926-9936.

Podust, L. M., Poulos, T. L., and Waterman, M. R. (2001) Crystal structure of cytochrome P450 14alpha -sterol demethylase (CYP51) from *Mycobacterium tuberculosis* in complex with azole inhibitors. *Proceedings of the National Academy of Sciences USA*. **98**: 3068-3073.

Podust, L. M., Yermalitskaya, L. V., Lepesheva, G. I., Dalmasso, V. N., Podu, E. A., and Waterman, M. R. (2004) Estriol bound and ligand-free structures of sterol 14alpha-demethylase. *Structure.* **12**: 1937.

Podust, L. M., Yermalitskaya, L. V., Kim, Y., and Waterman, M. R. Crystal structure analysis of the C37L/C151T/C442A-triple mutant of CYP51 from *Mycobacterium tuberculosis*. To be published.

Porter, T. D. and Coon, M. J. (1991) Cytochrome P450: multiplicity of isoforms, substrates and catalytic and regulatory mechanisms. *Journal of Biological Chemistry.* **266**: 13469-13472.

Potterton, E., Briggs, P., Turkenburg, M., and Dodson, E. (2003). A graphical user interface to the CCP4 program suite. *Acta Crystallographica D.* **59:** 1131-1137.

Poulos, T. L., Finzel, B. C., and Howard, A. J. (1987) High resolution crystal structure of cytochrome P-450cam. *Journal of Molecular Biology.* **195**: 687-700.

Poulos, T. L., Cupp-Vickery, J., and Li, H. "Structural Studies of Prokaryotic Cytochrome P450s" in Cytochrome P450 Structure, Mechanism, and Biochemistry (2nd edition). Edited by Ortiz de Montellano, P. R. Plenum Press, New York, USA. 1995.

Poulos, T. L., and Johnson, E. F. "Structures of Cytochrome P450 Enzymes" in Cytochrome P450 Structure, Mechanism, and Biochemistry (3rd edition). Edited by Ortiz de Montellano, P. R. Plenum Press, New York, USA. 2005.

Pouwels, P. H. "Survey of Cloning Vectors for *Escherichia coli*" in Essential Molecular Biology: A Practical Approach. Volume 1. 179 – 239. Edited by Brown, T. A. IRL Press, Oxford, UK. 1992.

Powell, H. (2006) Theory of Data Collection (online teaching resource). The Medical Research Council Laboratory of Molecular Biology.
http://www.mrc-lmb.cam.ac.uk/harry/lmbtalk/geometry.pdf#search=%22fully%20 recorded%20diffraction%20spots%20definition%22

Provencher, S. W. and Glockner, J. (1981) Estimation of globular protein secondary structure from circular dichroism. *Biochemistry.* **20**: 33-37.

Putnam, C. D., Arvai, A. S., Bourne, Y., and Tainer, J. A. (2000) Active and inhibited human catalase structures: Ligand and NADPH binding and catalytic mechanism. *Journal of Molecular Biology.* **296**: 295-309.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Research.* **33**: W116-W120.

Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963) Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology.* **7**: 955.

Ramachandran, G. and Gurumurthy, P. (2002) Effect of rifampicin & isoniazid on cytochrome P-450 in mycobacteria. *Indian Journal of Medical Research.* **116**: 140-144.

Raner, G. M., Hatchell, J. A., Dixon, M. U., Joy, T. K., Haddy, A. E., and Johnston, E. R. (2002) Regioselective peroxo-dependent heme alkylation in P450(BM3)-F87G by aromatic aldehydes: effects of alkylation on cataysis. *Biochemistry.* **41**(30): 9601-9610.

Rapoport, T. A., Hohne, W. E., Reich, J. G., Heitmann, P., and Rapoport, S. M. (1972) A kinetic model for the action of the inorganic pyrophosphatase from bakers' yeast. The activating influence of magnesium ions. *European Journal of Biochemistry.* **26**: 237-246.

Ravichandran, K. G., Boddupalli, S. S., Hasermann, C. A., Peterson, J. A., and Deisenhofer, J. (1993) Crystal structure of hemoprotein domain of P450BM-3, a prototype for microsomal P450s. *Science.* **6**: 731-736.

Reynolds, J. A and Tanford C. (1970) Binding of dodecyl sulfate to proteins at high binding ratios. Possible implications for the state of proteins in biological membranes. *Proceedings of the National Academy of Sciences.* **66**: 1002-1007.

Rigby-Duncan, K. E. and Stillman, M. J. (2006) Metal-dependent protein folding: Metallation of metallothionein. *Journal of Inorganic Biochemistry.* **19**: E-publication ahead of print.

Ridlington, J. W., and Butler, L. G. (1972) Yeast inorganic pyrophosphatase. I. Binding of pyrophosphate, metal ion, and metal ion-pyrophosphate complexes. *Journal of Biological Chemistry.* **247**: 7303-7307.

Roberts, B. E., and Paterson, B. M. (1973) Efficient translation of tobacco mosaic virus RNA and rabbit globin 9S RNA in a cell-free system from commercial wheat germ. *Proceedings of the National Academy of Science.* **70**: 2330-2334.

Roberts, E. S., Hopkins, N. E., Alworth, W. L., and Hollenberg, P. F. (1993) Mechanism-based inactivation of cytochrome P450 2B1 by 2-ethynylnaphthalene: Identification of an active-site peptide. *Chemical Research in Toxicology.* **6**: 470-479.

Rosenberg, A. (1996) Innovations, Novagen. 6:1-6.
www.emdbiosciences.com

Rossmann, M. G., and Blow, D. M. (1962) The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallographica.* **15**: 24-31.

Rost, B., Yachdav, G., and Liu, J. (2003) The PredictProtein Server. *Nucleic Acids Research 32(Web Server issue).* W321-W326.

Rupasinghe, S., Schuler, M. A., Kagawa, N., Yuan, H., Lei, L., Zhao, B., Kelly, S. L., Waterman, M. R., and Lamb, D. C. (2006) The cytochrome P450 gene family CYP157 does not contain EXXR in the K-helix reducing the absolute conserved P450 residues to a single cysteine. *FEBS Letters.* **580**: 6338-6342.

Sacchettini, J. C., Ronning, D. R., Klabunde, T., Besra, G. S., Vissa, V. D., Belisle, J. T. (2000) Crystal structure of the secreted form of antigen 85C reveals potential targets for mycobacterial drugs and vaccines. *Nature Structural Biology.* **7**: 141-146.

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., and Arnheim, N. (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anaemia. *Science.* **230**: 1350-1354.

Salminen, T., Kapyla, J., Heikinheimo, P., Kankare, J., Goldman, A., Heinonen, J., Baykov, A. A., Cooperman, B. S., and Lahti, R. (1995) Structure and function analysis of *Escherichia coli* inorganic pyrophosphatase: Is a hydroxide ion the key to catalysis? *Biochemistry.* **34**: 782-791.

Salusbury, T. "Clarification and extraction" in Protein Purification Methods: A Practical Approach. 86 - 95. Edited by Harris, E. L. V. and Angal, S. IRL Press, Oxford, UK. 1989.

Sambrook, J., and Russell, D. W. Molecular Cloning 3. 15.44-15.48. Cold Spring Harbour Laboratory Press, New York, USA. 2001.

Samygina, V. R., Antonyuk, S. V., Lamzin, V. S., and Popov, A.N. (2000) Improving the X-ray resolution by reversible flash-cooling combined with concentration screening, as exemplified with PPase. *Acta Crystallographica D.* **56**: 595-603.

Samygina, V. R., Popov, A. N., Rodina, E. V., Vorobyeva, N. N., Lamzin, V. S., Polyakov, K. M., Kurilova, S. A., Nazarova, T. I., and Avaeva, S. M. (2001). The structures of *Escherichia coli* inorganic pyrophosphatase complexed with Ca2+ or CaPPi at atomic resolution and their mechanistic implications. *Journal of Molecular Biology.* **314**: 633-645.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977) Chain termination procedure sequence-specific termination of an *in vitro* DNA synthesis reaction using modified nucleotide substrates. *Proceedings of the National Academy of Science.* **74**: 5463-5467.

Sassetti, C. M., Boyd, D. H., and Rubin, E. J. (2001) Comprehensive identification of conditionally essential genes in mycobacteria. *Proceedings of the National Academy of Sciences.* **98**(22): 12712-12717.

Sassetti, C. M and Rubin, E. J. (2003) Genetic requirements for mycobacterial survival during infection. *Proceedings of the National Academy of Sciences.* **100**(22): 12989-12994.

Schachtschabel, D., und Zillig, W. (1959) Untersuchungen zur biosynthese der proteine. I. Uber den einbau C14-markierter aminosauren ins protein zellfreier nucleoproteid-enzyme-systeme aus *E. coli* B. *Hoppe-Seyler's Z. Physiol. Chem.* **314**: 262-275.

Schreier, E. (1980). Reversible acid dissociation of thermostable inorganic pyrophosphatase from *Bacillus stearothermophilus. FEBS Letters.* **109**: 67-70.

Scott, E. E., White, M. A., He, Y. A., Johnson, E. F., Stout, C. D., and Halpert, J. R. (2004) Structure of mammalian cytochrome P450 2B4 complexed with 4-(4-chlorophenyl)imidazole at 1.9 Å resolution: Insight into the range of P450 conformations and coordination of redox partner binding. *Journal of Biological Chemistry.* **279**: 27294-27301.

Seidel, H.M. (1992). Phosphonate biosynthesis: molecular cloning of the gene for phosphoenolpyruvate mutase from *Tetrahymena pyriformis* and overexpression of the gene product in *Escherichia coli. Biochemistry.* **31**: 2598–2608.

Seward, H., Roujeinikova, A., McLean, K. J., Munro, A. W., and Leys, D. Novel azole ligation in CYP121. To be Published.

Shaik, S. and De Visser, S. P. "Cytochrome P450s in Plants" in Cytochrome P450 Structure, Function, and Mechanism (3rd edition). Edited by Ortiz de Montellano, P. R. Plenum Press, New York, USA. 2005.

Sherman, D. H., Li, S., Yermalitskaya, L. V., Kim, Y., Smith, J. A., Waterman, M. R., and Podust, L. M. (2006) The structural basis for substrate anchoring, active site selectivity, and product formation by P450 Pikc from *Streptomyces venezuelae. Journal of Biological Chemistry.* **281**: 26289.

Shimizu, H., Park, S., Lee, D., Shoun, H., and Shiro, Y. (2000) Crystal structures of cytochrome P450nor and its mutants (Ser286-->Val, Thr) in the ferric resting state at cryogenic temperature: A comparative analysis with monooxygenase cytochrome P450s. *Journal of Inorganic Biochemistry*. **81**: 191-205.

Shintani, T., Hachimori, A. (1998) Cloning and expression of a unique inorganic pyrophosphatase from *Bacillus subtilis*. *FEBS Letters*. *439*: 263–266.

Shuman, S. (1994) Novel approach to molecular cloning and polynucleotide synthesis using vaccinia DNA topoisomerase. *Journal of Biological Chemistry*. **269**: 32678-32684.

Sivula, T., Salminen, A., Parfenyev, A. N., Pohjanjoki, P., Goldman, A., Cooperman, B. S., Baykov, A. A., and Lahti, R. (1999) Evolutionary aspects of inorganic pyrophosphatase. *FEBS Letters*. **454**: 75-80.

Sligar, S. G. (1976) Coupling of spin, substrate, and redox equilibria in cytochrome P-450. *Biochemistry*. **15**: 5399-5406.

Sligar, S. G. and Gunsalus, I. C. (1976) A thermodynamic model of regulation: Modulation of redox equilibria in camphor monoxygenase. *Proceedings of the National Academy of Sciences*. **73**: 1078-1082.

Smith, F. A. Handbook of Experimental Pharmacology. Volume 2, parts 1 and 2. Springer-Verlag, New York, USA. 1970.

Smith, C. V., Huang, C., Miczak, A., Russell, D. G., Sacchettini, J. C., Honer, K. (2003) Biochemical and structural studies of malate synthase from *Mycobacterium tuberculosis*. *The Journal of Biological Chemistry*. **278**(3): 1735-1743.

Sorensen, H. P. and Mortensen, K. K. (2005) Advanced genetic strategies for recombinant protein expression in *Escherichia coli*. *Journal of Biotechnology*. **15**: 113-128.

Sreerama, N. and Woody, R. W. (2000) Estimation of protein secondary structure from CD spectra: Inclusion of denatured proteins with native protein in the analysis. *Analytical Biochemistry*. **287**: 252-260.

Sreerema, N. and Woody, R. W. (1993) A self-consistent method for the analysis of protein secondary structure from circular dichroism. *Analytical Biochemistry*. **209**: 32-44.

Sreerema, N., Venyaminov, S. Y., and Woody, R. W. (1999) Estimation of the number of helical and strand segments in proteins using CD spectroscopy. *Protein Science*. **8**: 370-380.

Stout, G. H. and Jensen, L. H. X-ray Structure Determination: A Practical Guide (2nd edition). John Wiley and Sons Inc., New York, USA. 1989.

Studier, F. W. and Moffatt, B. A. (1986) Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *Journal of Molecular Bioloy*. **189**: 113-130.

Szczesna-Skorupa, E., Straub, P., and Kemper, B. (1993) Deletion of a conserved tetrapeptide, PPGP, in P450 2C2 results in a loss of enzymatic activity without a change in its cellular location. *Archives of Biochemistry and Biophysics*. **304**: 170-75.

Szczesna-Skorupa, E., Ahn, K., Chen, C-D., Doray, B., and Kemper, B. (1995) The Cytoplasmic and N-terminal Transmembrane Domains of Cytochrome P450 Contain Independent Signals for Retention in the Endoplsmic Reticulum. *Journal of Biological Chemistry*. **270**: 24327-33.

Tammenkoski, M., Benini, S., Magretova, N. N., Baykov, A. A., and Lahti, R. (2005) An unusual his-dependant family 1 pyrophosphatase from *Mycobacterium tuberculosis*. *Journal of Biological Chemistry*. **280**: 41819-41826.

Teplyakov, A., Obmolova, G., Wilson, K. S., Ishii, K., Kaji, H., Samejima, T., and Kuranova, I. (1994). Crystal structure of inorganic pyrophosphatase from *Thermus thermophilus*. *Protein Science*. **3**: 1098-1107.

Terwilliger, T. C., Park, M. S., Waldo, G. S., Berendzen, J., Hung, L. W., Kim, C. Y., Smith, C. V., Sacchettini, J. C., Bellinzoni, M., Bossi, R., De Rossi, E., Mattevi, A., and Rupp, B. *et al.* (2003) The TB structural genomics consortium: a resource for *Mycobacterium tuberculosis* biology. *Tuberculosis.* **83**: 223-249.

Terzyan, S. S., Voronova, A. A., Smirnova, E. A., Kuranova, I. P., Nekrasov, Yu. V., Arutyunyan, E. G., Vainshtein, B. K., Hohne, W., and Hansen, G. (1984) Spatial structure of inorganic pyrophosphatase from yeast at 3 Å resolution. *Bioorganicheskaya Khimiya.* **10**: 1469-1482.

Testa, B. and Jenner, P. (1981) Inhibitors of Cytochrome P-450s and their mechanisms of action. *Drug Metabolism Reviews.* **12**: 111-117.

Thompson, J.D., Higgins, D.G., Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research.* **22**: 4673-4680.

Thompson, J.D., Plewnial, F., Thierry, J-C. and Poch, O. (2000) Rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acid Research.* **28**(15): 2919-2926.

Thorne, H. V. (1966) Electrophoretic separation of polyoma virus DNA from host cell DNA. *Virology.* **29**: 234-239.

Timasheff, S. N. and Arakawa, T. A Practical Approach in Protein Structure. Edited by Creighton, T. E. IRL Press, Oxford, UK. 1989.

Tissiéres, A., Schlessinger, D., and Gros, F. (1960) Amino acid incorporation into proteins by *E. coli* ribosomes. *Proceedings of the National Academy of Science.* **46**: 1450-1463.

Tosha, T., Yoshioka, S., Takahashi, S., Ishimori, K., Shimada, H. and Morishima, I. (2003) NMR Study on the Structural Changes of Cytochrome P450cam upon the Complex Formation with Putidaredoxin. *Journal of Biological Chemistry.* **278**(41): 39809-39821.

Triccas, J. A., and Gicquel, B. (2001) Analysis of stress- and host cell-induced expression of the *Mycobacterium tuberculosis* inorganic pyrophosphatase. *BMC Microbiology.* **1**: 3 electronic publication.

Tsai, R., Yu, C. A., Gunsalus, I. C., Peisach, J., Blumberg, W., Orme-Johnson, W. H., and Beinert, H. (1970) Spin-state changes in cytochrome P-450$_{cam}$ on binding of specific substrates. *Proceedings of the National Academy of Sciences.* **66**: 1157-1163.

Tullius, M. V. (2001) High extracellular levels of *Mycobacterium tuberculosis* glutamine synthetase and superoxide dismutase in actively growing cultures are due to high expression and extracellular stability rather than to a protein-specific export mechanism. *Infectious Immunology.* **69**: 6348-6363.

Tyson, C. A., Lipscomb, J. D., and Gunsalus, I. C. (1972) The role of putidaredoxin and P-450cam in methylene hydroxylation. *Journal of Biological Chemistry.* **247**: 5777-5784.

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

Van Etten, R. L., Davidson, R., Stevis, P. E., MacArthur, H., and Moore, D. L. (1991) Covalent structure, disulfide bonding, and identification of reactive surface and active site residues of human prostatic acid phosphatase. *Journal of Biological Chemistry.* **266**: 2313-2319.

Van Montfort, R. L. M., Congreve, M., Tisi, D., Carr R., and Jhoti, H. (2003) Oxidation state of the active-site cysteine in protein tyrosine phosphatase 1B. *Nature.* **423**: 773-777.

Van Stokkum, I. H. M., Spoelder, H. J. W., Bloemendal, M., Van Grondelle, R., and Groen, F. C. A. (1990) Estimation of protein secondary structure and error analysis from CD spectra. *Analytical Biochemistry.* **191**: 110-118.

Verras, A., Alian, A., and Montellano, P. R. (2006) Cytochrome P450 active site plasticity: Attenuation of imidazole binding in cytochrome P450cam by an L244A mutation. *Protein Engineering Design and Selection.* **19**: 491-496.

Vihinen, M., Lundin, M., and Baltscheffsky, H. (1992) Computer modeling of two inorganic pyrophosphatases. *Biochemical and Biophysical Research Communications.* **186**: 122-128.

Voet, D., Voet, J. G., and Pratt, C. W. Fundamentals of Biochemistry. 99-103.
John Wiley Press, New York, USA. 1999.

Vriend, G. (1990) WHATIF: A molecular modelling and drug design program. *Journal of Molecular Graphics and Modelling.* **8**: 52-56.

Wachenfeldt, C. V. and Johnson, E. J. "Structures of Eukaryotic Cytochrome P450 Enzymes" in Cytochrome P450 Structure, Mechanism, and Biochemistry (2nd edition). Edited by Ortiz de Montellano, P. R. Plenum Press, New York, USA. 1995.

Walker, J. M. The Protein Protocols. Humana Press, Totowa, USA. 1998.

Wanatabe, Y. and Groves, J. T. "Molecular Mechanism of Oxygen Activation by Cytochrome P-450" in The Enzymes (3rd edition). Volume XX. Edited by Sigman, D. Academic Press, New York, USA. 1992.

Wang, J. and Boisvert, D.C. (2003) Structural basis for GroEL-assisted protein folding from the crystal structure of (GroEL-KMgATP)14 at 2.0A resolution. *Journal of Molecular Biology.* **327**: 843-855.

Watson, J. D. and Crick, F. H. C. (1953) A structure for deoxyribose nucleic acid. *Nature.* **171**: 737-738.

Werck-Reichhart, D. and Feyereisen, R. (2000) Cytochromes P450: A success story. *Genome Biology.* **1**(6): 3003.1-3003.9.

Werck-Reichhart, D., Bak, S., and Paquette. S. "Cytochromes P450" in The Aribidopsis Book. Edited by Somervile, C. R. and Meyerowitz, E. M. American Society of Pland Biologists, Rockville, USA. 2002.

Whittington, P. N. "Clarification and Extraction" in Protein Purification Methods: A Practical Approach. 67 - 85. Edited by Harris, E. L. V. and Angal, S. IRL Press, Oxford, UK. 1989.

WHO (2006) Tuberculosis FACTS 2006. http://www.who.int/tb/publications/2006/tb_factsheet_2006_1_en.pdf

Williams, P. A., Cosme, J., Sridhar, V., Johnson, E. F., and McRee, D. E. (2000) Mammalian microsomal cytochrome P450 monooxygenase: Structural adaptations for membrane binding and functional diversity. *Molecular Cell.* **5**: 121-131.

Williams, P. A., Cosme, J., Ward, A., Angove, H. C., Matak Vinkovic, D., and Jhoti, H. (2003) Crystal structure of human cytochrome P450 2C9 with bound warfarin. *Nature.* **424**: 464.

Williams, P. A., Cosme, J., Vinkovic, D. M., Ward, A., Angove, H. C., Day, P. J., Vonrhein, C., Tickle, I. J., and Jhoti, H. (2004) Crystal structures of human cytochrome P450 3A4 bound to metyrapone and progesterone. *Science.* **305**: 683.

Wilson, A. J. C. (1949) The probability distribution of X-ray intensities. *Acta Crystallographica.* **2**: 318-321.

Wink, D. A., Osawa, Y., Darbyshe, J. F., Jones, C. R., Eshenaur, S. C., and Nims, R. W. (1993) Inhibition of cytochromes P450 by nitric oxide and a nitric oxide-releasing agent. *Archives of Biochemistry and Biophysics.* **300**: 115-123.

Winn, M. D., Isupov, M. N., and Murshudov, G. N. (2001) Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta Crystallographica D.* **57**: 122-123.

Wu, C. W., Terkeltaub, R., and Kalunian, K. C. (2005) Calcium-containing crystals and osteoarthritis: implications for the clinician. *Current Rheumatology Reports.* **7**(3): 213-219.

Yamazaki, S. (1993) Importance of the proline-rich region following signal-anchor sequence in the formation of correct conformation of microsomal cytochrome P-450. *Journal of Biochemistry*. **114**: 652-57.

Yokoyama, S. (2003) Protein expression systems for structural genomics and proteomics. *Current Opinion in Chemical Biology*. **7**(1): 39-43.

Yano, J. K., Koo, L. S., Schuller, D. J., Li, H., Ortiz de Montellano, P. R., and Poulos, T.L. (2000) Crystal structure of a thermophilic cytochrome P450 from the archaeon *Sulfolobus solfataricus*. *Journal of Biological Chemistry*. **275**: 31086-31092.

Yoshikawa, K-I. and Go, M. (1992) Hydrogen bond network of cytochrome P-450cam: A network connecting the haem group with helix K. *Biochimica et Biophysica Acta*. **1122**: 41-44.

Yoshioka, S., Takahashi, S., Hori, H., Ishimori, K., and Morishima, I. (2001) Proximal cysteine residue is essential for the enzymatic activities of cytochrome P450cam. *European Journal of Biochemistry*. **268**: 252-259.

Young, T.W., Kuhn, N.J., Wadeson, A., Ward, S., Burges, D., Cooke, G.D. (1998) *Bacillus subtilis* ORF yybQ encodes a manganese-dependent inorganic pyrophosphatase with distinctive properties: the first of a new class of soluble pyrophosphatase? *Microbiology*. **144**: 2563–2571.

Yuen, L. K., Leslie, D., and Coloe, P. J. (1999) Bacteriological and molecular analysis of rifampin-resistance *Mycobacterium tuberculosis* strains isolated in Australia. *Journal of Clinical Microbiology*. **37**: 3844-3850.

Yun, C-H., Song, M., Ahn, T., and Kim, H. (1996) Conformation change of cytochrome P450 1A2 induced by sodium chloride. *The Journal of Biological Chemistry*. **271**: 31312-31316.

Yun, C-H., Ahn, T., and Guengerich, F. P. (1998) Conformational change and activation of cytochrome P450 2B1 induced by salt and phospholipid. *Archives of Biochemistry and Biophysics*. **356**: 229-238.

Zerbe, K., Pylypenko, O., Vitali, F., Zhang, W., Rouset, S., Heck, M., Vrijbloed, J. W., Bischoff, D., Bister, B., Sussmuth, R. D., Pelzer, S., Wohlleben, W., Robinson, J. A., and Schlichting, I. (2002) Crystal structure of OxyB, a cytochrome P450 implicated in an oxidative phenol coupling reaction during vancomycin biosynthesis. *Journal of Biological Chemistry*. **277**: 47476-47485.

Zhang, W., Ramamoorthy, Y., Kilicarslan, T., Nolte, H., Tyndale, R. F., and Sellers, E. M. (2002) Inhibition of cytochromes P450 by antifungal imidazole derivatives. *Drug Metabolism and Disposition*. **30**(3): 314-318.

Zhao, B., Guengerich, F. P., Voehler, M., and Waterman, M. R. (2005) Role of active site water molecules and substrate hydroxyl groups in oxygen activation by cytochrome P450 158A2: A new mechanism of proton transfer. *Journal of Biological Chemistry*. **280**: 42188-42197.

Zhou, S., Yung Chan, S., Cher Goh, B., Chan, E., Duan, W., Huang, M., and McLeod, H. L. (2005) Mechanism-based inhibition of cytochrome P450 3A4 by therapeutic drugs. *Clinical Pharmacokinetics*. **44**(3): 279-304.

Zhou, W., Tempel, W., Liu, Z.-J., Chen, L., Clancy Kelley, L.-L., Dillard, B. D., Hopkins, R. C., Arendall III, W. B., Rose, J. P., Eneh, J. C., Hopkins, R. C., Jenney Jr., F. E., Lee, H. S., Li, T., Poole II, F. L., Shah, C., Sugar, F. J., Adams, M. W. W., Richardson, J. S., Richardson, D. C., Wang, B.-C. Inorganic pyrophosphatase from *Pyrococcus furiosus*. To be published.

Zyryanov, A. B., Tammenkoski, M., Salminen, A., Kolomiytseva, G. Y., Fabrichniy, I. P., Goldman, A., Lahti, R., and Baykov, A. A. (2004) Site-specific effects of zinc on the activity of family II pyrophosphatase . *Biochemistry*. **43**(45): 14395-14402.

Appendix 1 – Additional gel electrophoresis photographs from cell-free expression

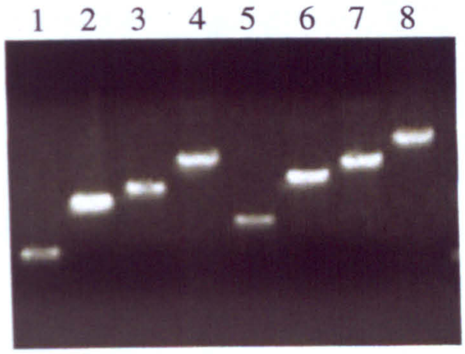


Figure A1: Examples of gene targets successfully amplified by 2-step PCR, visualised by 1 % agarose gel electrophoresis (section 4.2.3a). 1st and 2nd PCR products for: **Rv2718**, lanes 1 – 2; **Rv2776c**, 3 – 4; **Rv2986c**, 5 – 6; **Rv3042c**, 7 – 8.

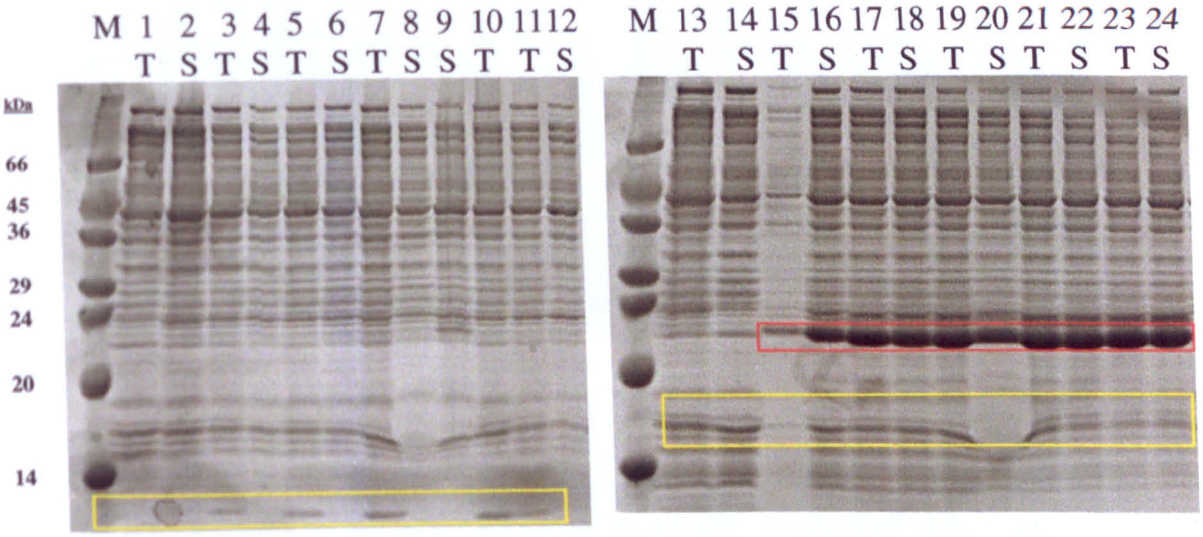

M 1 2 3 4 5 6 7 8 9 10 11 12
T S T S T S T S S T T S
M 13 14 15 16 17 18 19 20 21 22 23 24
T S T S T S T S T S T S
kDa
66
45
36
29
24
20
14

Figure A2: 10 % SDS-PAGE from cell-free time-course expression study of targets which were not expressed correctly (section 4.2.3c). Total extracts (T) and soluble fractions (S) from small-scale reactions incubated at 37 °C over 0 to 24 hours. **Rv0359: 0 hours,** lanes 1 - 2; **2 hours,** 3 - 4; **4 hours,** 5 - 6; **6 hours,** 7 - 8; **8 hours,** 9 - 10; **24 hours,** 11 - 12. **Rv2718: 0 hours,** lanes 13 - 14; **2 hours,** 15 - 16; **4 hours,** 17 - 18; **6 hours,** 19 - 20; **8 hours,** 21 - 22; **24 hours,** 23 - 24. Molecular weight markers in lane M. Target proteins highlighted in red and unknown contaminants in yellow.

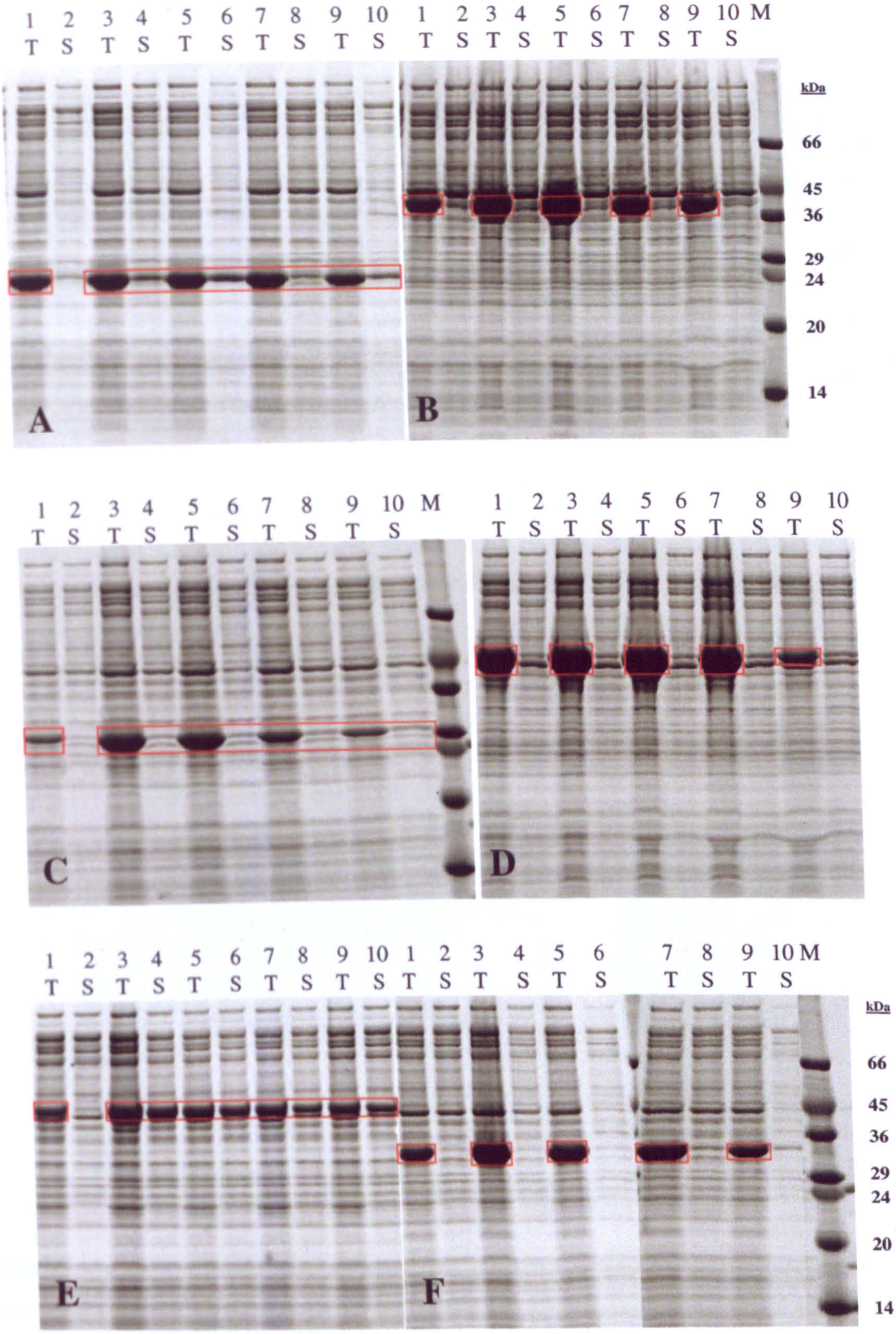

1 2 3 4 5 6 7 8 9 10 1 2 3 4 5 6 7 8 9 10 M
T S T S T S T S T S T S T S T S T S T S
kDa
66
45
36
29
24
20
14
A
B
1 2 3 4 5 6 7 8 9 10 M
T S T S T S T S T S
1 2 3 4 5 6 7 8 9 10
T S T S T S T S T S
C
D
1 2 3 4 5 6 7 8 9 10 1 2 3 4 5 6 7 8 9 10 M
T S T S T S T S T S T S T S T S T S T S
kDa
66
45
36
29
24
20
14
E
F

Figure A3: 10 % SDS-PAGE from optimisation of cell-free expression conditions by the addition of detergents (section 4.2.5c). Total extracts (T) and soluble fractions (S) from small-scale reactions incubated at 30 °C for 4 hours. **(A)** Rv0185, **(B)** Rv2776c, **(C)** Rv3717, **(D)** Rv3915, **(E)** Rv2388c, and **(F)** Rv0247c: **No detergent** in lanes 1 – 2; **Brij-35**, 3 - 4 (0.5 %) and 5 - 6 (1 %); **Digitonin**, 7 - 8 (0.5 %) and 9 - 10 (1 %). Molecular weight marker in lane M. Target proteins highlighted in red.

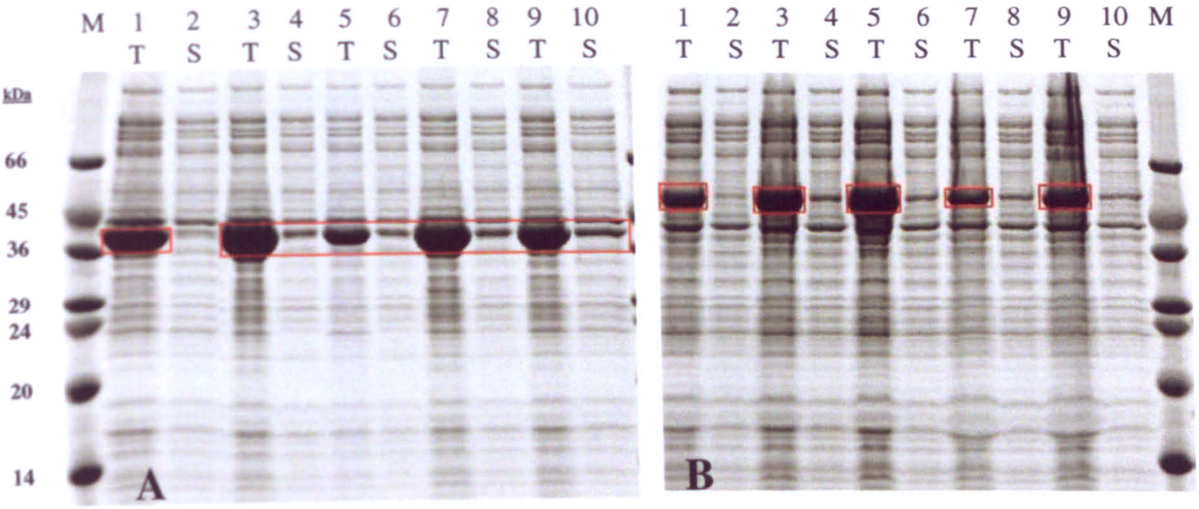


Figure A4: 10 % SDS-PAGE from optimisation of cell-free expression conditions by the addition of detergents (section 4.2.5c). Total extracts (T) and soluble fractions (S) from small-scale reactions incubated at 30 °C for 4 hours. **(A)** Rv3534c and **(B)** Rv3545c: **No detergent** in lanes 1 – 2; **Brij-35**, 3 - 4 (0.5 %) and 5 - 6 (1 %); **Digitonin**, 7 - 8 (0.5 %) and 9 - 10 (1 %). Molecular weight marker in lane M. Target proteins highlighted in red.

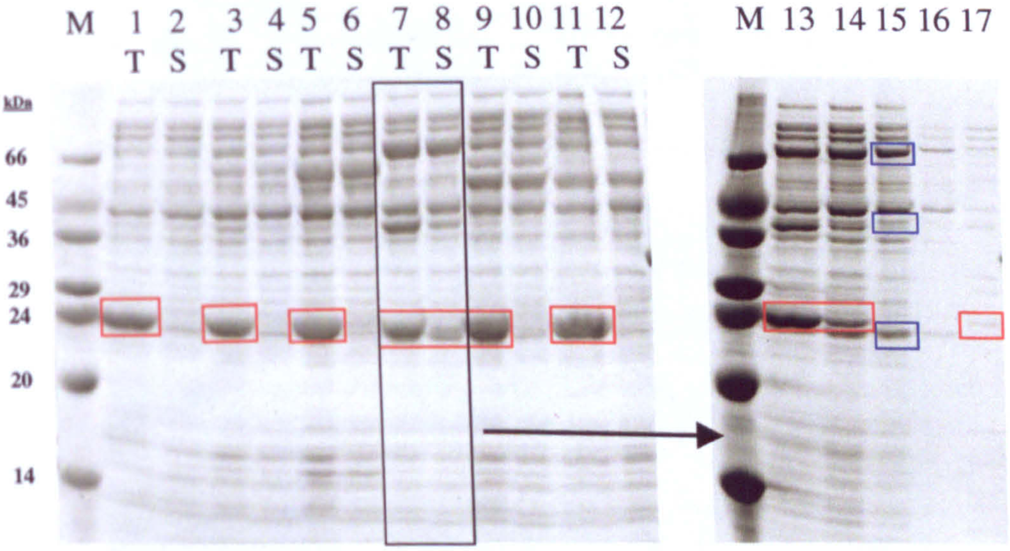


Figure A5: 10 % SDS-PAGE from optimisation of **Rv0185** expression conditions by the addition of molecular chaperones (section 4.2.5c). Total extracts (T) and soluble fractions (S) from small-scale reactions incubated at 30° C for 4 hours. **No chaperones** in lanes 1 – 2; with **dnaJ-dnaK-grpE-groEL-groES**, 3 – 4; **groEL-groES**, 5 – 6; **dnaJ-dnaK-grpE**, 7 – 8; **groEL-groES-trigger factor**, 9 – 10; **trigger factor**, 11 - 12. Fractions from affinity chromatography of **Rv0185** synthesised in the presence of dnaJ-dnaK-grpE: **Total**, 13; **Soluble**, 14; **Flow-through**, 15; **Wash**, 16; **Elution**, 17. Molecular weight markers in lane M. Rv0185 highlighted in red and chaperones in blue.

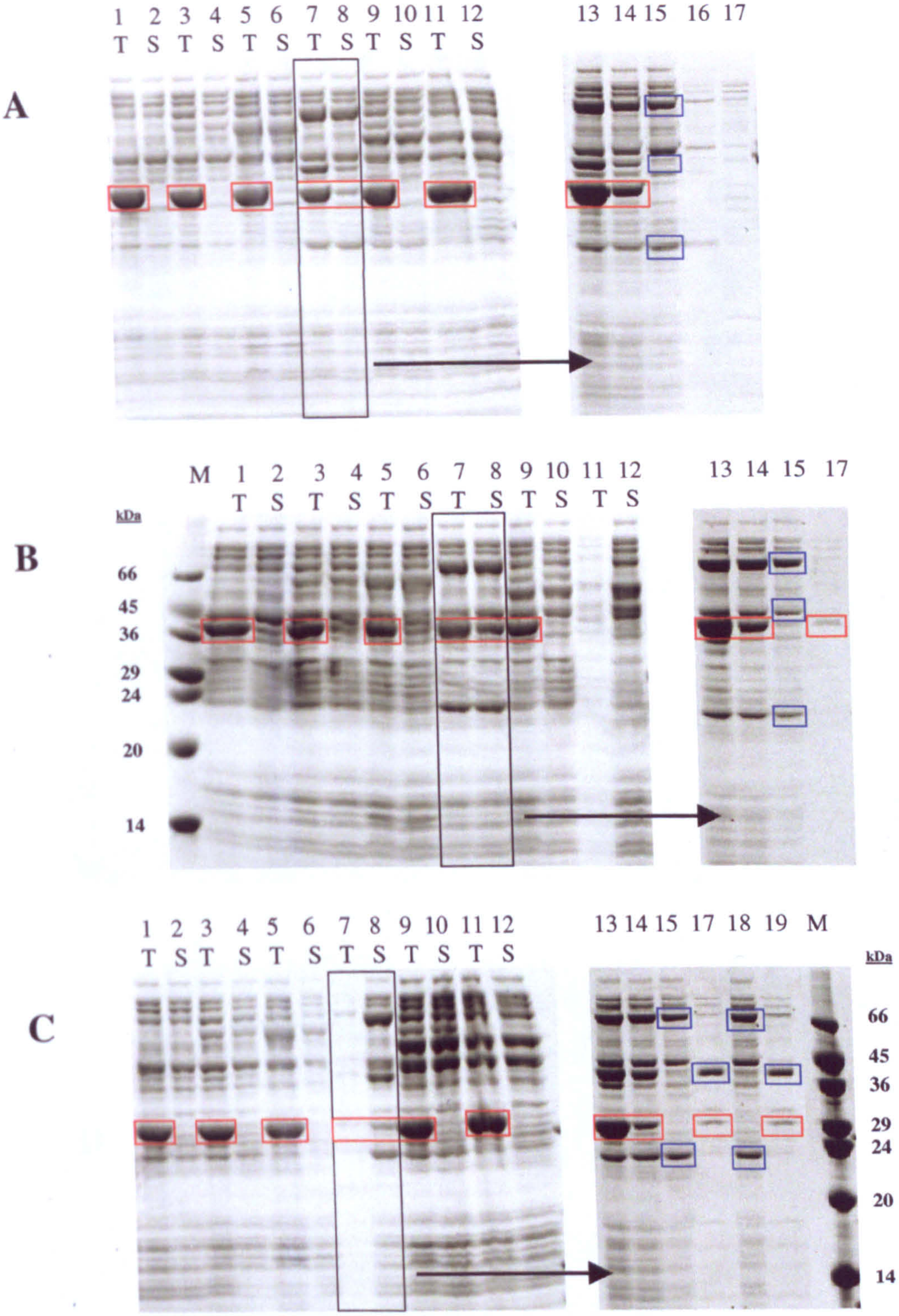


Figure A6: 10 % SDS-PAGE from optimisation of expression conditions by the addition of molecular chaperones (section 4.2.5c). Total extracts (T) and soluble fractions (S) from small-scale reactions incubated at 30° C for 4 hours. **(A)** Rv0247c, **(B)** Rv2776c, and **(C)** Rv3717: **No chaperones** in lanes 1 – 2; with **dnaJ-dnaK-grpE-groEL-groES**, 3 – 4; **groEL-groES**, 5 – 6; **dnaJ-dnaK-grpE**, 7 – 8; **groEL-groES-trigger factor**, 9 – 10; **trigger factor**, 11 - 12. Fractions from affinity chromatography of target protein synthesised in the presence of dnaJ-dnaK-grpE: **Total**, 13; **Soluble**, 14; **Flow-through**, 15; **Wash**, 16; **Elution**, 17; **Flow-through after ATP incubation**, 18; **Elution after ATP incubation**, 19. Molecular weight markers in lane M. Rv0185 highlighted in red and chaperones in blue.

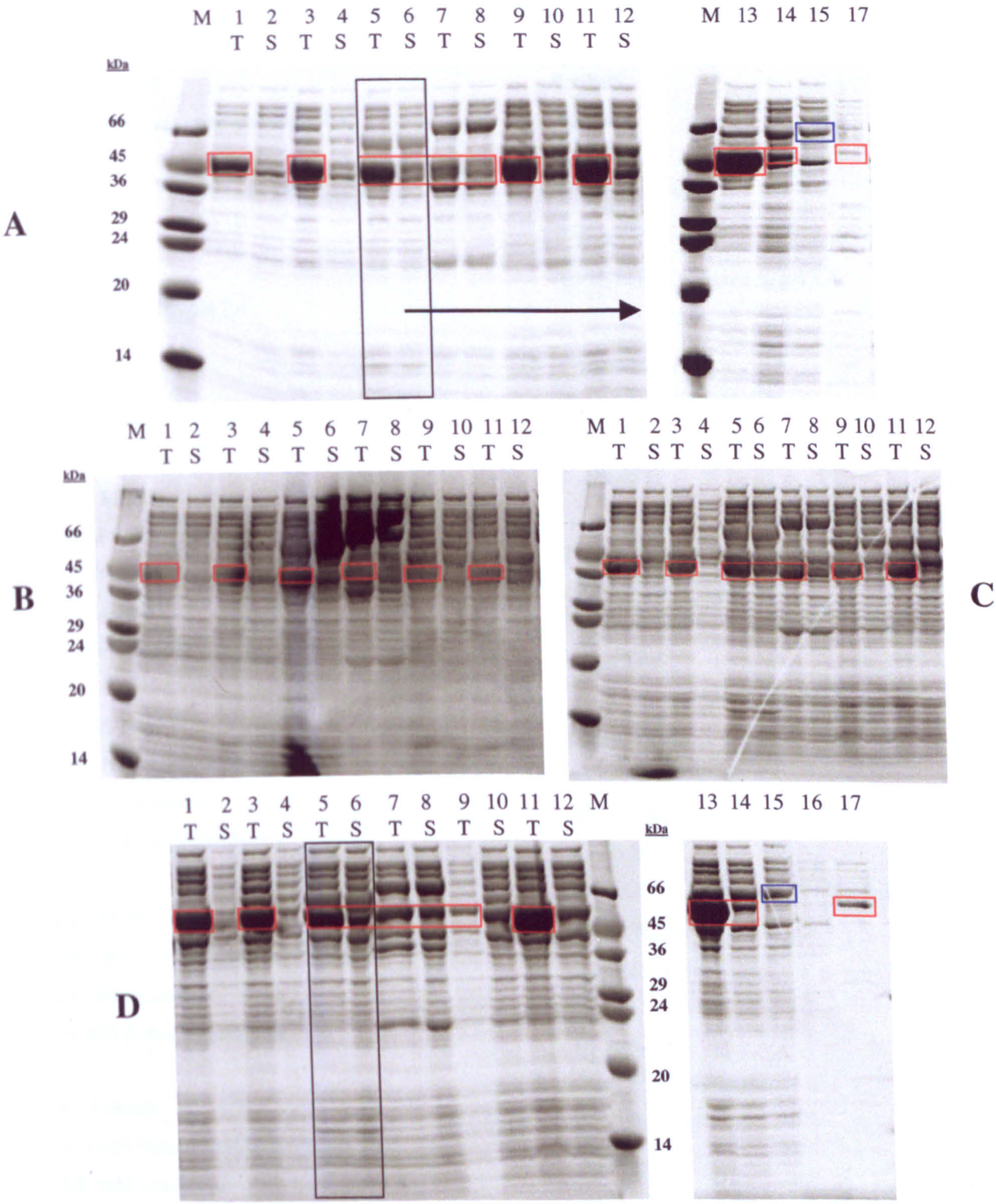


Figure A7: 10 % SDS-PAGE from optimisation of expression conditions by the addition of molecular chaperones (section 4.2.5c). Total extracts (T) and soluble fractions (S) from small-scale reactions incubated at 30° C for 4 hours. **(A)** Rv3915, **(B)** Rv2388c, **(C)** Rv3534c, and **(D)** Rv3545c: **No chaperones** in lanes 1 – 2; with **dnaJ-dnaK-grpE-groEL-groES**, 3 – 4; **groEL-groES**, 5 – 6; **dnaJ-dnaK-grpE**, 7 – 8; **groEL-groES-trigger factor**, 9 – 10; **trigger factor**, 11 - 12. Fractions from affinity chromatography of target protein synthesised in the presence of dnaJ-dnaK-grpE: **Total**, 13; **Soluble**, 14; **Flow-through**, 15; **Wash**, 16; **Elution**, 17. Molecular weight markers in lane M. Target proteins highlighted in red and chaperones in blue.

Appendix 2 - Buffer and media compositions

1 x Rv125-GF

50 mM Potassium phosphate pH 7.4
500 mM KCl
1 mM dTT

1 x 125-K

50 mM Tris-HCl pH 7.4

1 x 125-Lysis

500 mM Potassium phosphate pH 7.4
10 mM Imidazole
5 mM $MgSO_4$
10 mM β-mercaptoethanol
1/1000 diluted protease inhibitor complex III (Calbiochem)
1 µg/ml DNAseI

1 x 125-NiA

500 mM Potassium phosphate pH 7.4
10 mM Imidazole
10 mM β-mercaptoethanol

1 x 125-NiB

500 mM Potassium phosphate pH 7.4
20 mM Imidazole
10 mM β-mercaptoethanol

1 x 125-NiC

500 mM Potassium phosphate pH 7.4
300 mM Imidazole
10 mM β-mercaptoethanol

1 x 125-Ni-I

500 mM Potassium phosphate pH 7.4

1 x PBS

8 g NaCl
0.2 g KCl
1.44 g Na_2HPO_4
0.24 g KH_2PO_4

To 1L in H_20 adjusted to pH 7.4

1 x ppa-AxA

50 mM Tris-HCl pH 8.0
50 mM NaCl
1 mM dTT

1 x ppa-AxB

50 mM Tris-HCl pH 8.0
1 M NaCl
1 mM dTT

1 x ppa-Lysis

50 mM Tris-HCl pH 8.0
750 mM NaCl
1 mM dTT
1/1000 diluted protease inhibitor complex III (Calbiochem)
1 µg/ml DNAseI

1 x ppa-NiA

50 mM Tris-HCl pH 8.0
750 mM NaCl
1 mM dTT
1/1000 diluted protease inhibitor complex III (Calbiochem)

1 x ppa-NiB

50 mM Tris-HCl pH 8.0
300 mM NaCl
750 mM Imidazole
1 mM dTT

1 x 125-P

50 mM Tris-HCl pH 7.4

500 mM KCl

1 x 125-S

50 mM Potassium phosphate pH 7.4

1 x CF-A

50 mM NaH_2PO_4 pH 8.0

750 mM NaCl

1/1000 dilution Complete protease inhibitors (Roche)

1 mM DTT

1 x CF-B

50 mM NaH_2PO_4 pH 8.0

300 mM NaCl

500 mM Imidazole

1/1000 dilution Complete protease inhibitors (Roche)

1 mM DTT

1 x CF-C

50 mM NaH_2PO_4 pH 8.0

50 mM NaCl

1 mM dTT

1 x CF-D

50 mM NaH_2PO_4 pH 8.0

1 M NaCl

1 mM dTT

1 x CF-E

50 mM NaH_2PO_4 pH 8.0

150 mM NaCl

1 mM dTT

1 x CF-F

20 mM Tris-HCl pH 8.0

10 mM NaCl

1 mM dTT

1 x S30

10 mM Tris-acetate pH 8.2

60mM potassium acetate

SOC medium

2 % Tryptone

0.5 % Yeast extract

0.4 % glucose

10 mM NaCl

2.5 mM KCl

5 mM $MgCl_2$

5 mM $MgSO_4$

50 x TAE

242 g Tris base

57.1 ml Glacial acetic acid

18.6 g EDTA

To 1 L in H_2O

10 x TBE

108 g Tris base

55 g Boric acid

9.3 g Na_4EDTA

To 1 L in H_2O (end pH 8.3)

Terrific broth (TB)

12 g tryptone

24 g yeast extract

4 ml glycerol

To 900 ml in H_20

2.31 g KH_2PO_4 monobasic

12.54 g K_2HPO_4 dibasic

To 100ml in H_2O

Autoclave separately and add phosphate solution once cooled below 60 °C.

1 x CF-G

20 mM Tris-HCl pH 8.0
1 M NaCl
1 mM dTT

1 x CF-H

20 mM Tris-HCl pH 8.0
150 mM NaCl
1 mM dTT

6 x Loading dye solution

10 mM Tris-HCl pH 7.6
0.03 % bromophenol blue
0.03 % xylene cyanol FF
60 % glycerol
60m M EDTA

Luria-Bertani (LB) broth

10 g Bacto-tryptone
5 g Yeast extract
10 g NaCl
To 1 L in H_20. Autoclave.

1 x Na-Lysis

50 mM Sodium phosphate pH 8.0
750 mM NaCl
1 mM dTT
1/1000 diluted protease inhibitor complex III
(Calbiochem)
1 µg/ml DNAseI

10 x TES

108 g Tris base
55 g Boric acid
9.3 g Na_4 EDTA

To 1 L in water. Do not adjust pH (pH 8.3).

1 x TES

10 mM Tris-HCl pH 7.4
5 mM EDTA
1 % SDS
Autoclave.

10 x Tris-glycine running buffer

30.3 g Tris base
144 g Glycine
10 g SDS
To 1L in H_20