Biochemical and structural characterisation of a thermophilic Aldo-Keto Reductase from

Thermotoga maritima

Submitted by Simon Carl Willies, to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Biosciences, March 2009.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

.... (signature)

Abstract

The Aldo-Keto Reductases (AKR) are a group of oxidoreductase enzymes structurally and mechanistically distinct from the Alcohol Dehydrogenases (ADH). The AKRs are of importance for their ability to produce industrially useful compounds including chiral secondary alcohols. The ADH family have traditionally been exploited for chiral alcohol production; the AKR family have currently been underexploited for chiral alcohol production and present the opportunity to search for novel oxidoreductases with properties and substrate specificities distinct from the ADH enzymes.

The AKR studied here, from the hyperthermophilic bacteria *Thermotoga maritima* has been characterised with respect to its biochemical and structural properties, and its potential as a biocatalyst evaluated. This enzyme is the second example of a thermophilic AKR to have its three dimensional structure solved, the other also being from *Thermot. maritima*. The AKR studied exhibits high stability with respect to temperature and moderate amounts of organic solvents. A large preference for the reduction reaction compared to the oxidation reaction was found, which has previously been observed in other AKRs. The X-ray crystal structure was solved to 2.6Å resolution in the apo form. The final structure has three loop sections which were not located due to disorder within the crystal, which are expected to become ordered upon cofactor and substrate binding. A section of one of these missing loops was found to bind at the active site of the enzyme, with a glutamate occupying the site of substrate carbonyl binding. The formation of a dimer, increased helix-dipole stabilisation and long distance ion pair interactions all act to increase thermostability of the AKR with respect to its mesophilic homologues.

The X-ray crystal structure of *Escherichia coli* bacterioferritin has also been solved to 1.9Å resolution, which was co-purified along with the recombinant AKR enzyme. This structure shows the symmetrical binding of a heme molecule on the local two-fold axis between subunits and the binding of two metal atoms to each subunit at the ferroxidase centre. These metal atoms have been identified as zinc by the anaylsis of the structure and X-ray data and confirmed by microPIXE experiments. For the first time the heme has been shown to be linked to the internal and external environments via a cluster of waters positioned above the heme molecule. This information has provided a greater insight into the function and mechanism of bacterioferritin.

Acknowledgements

I would like to thank all those who have helped and supported me through the course of my PhD. Specific mention should be given to some. Misha Isupov, who has provided his help, knowledge and experience throughout all the stages of X-ray crystallography work, and without whom some of the problems encountered along the way might not have been solved. I would like to thank my supervisors Jenny Littlechild and Matt Bycroft who have provided helpful conversations and suggestions throughout the project. Thank you also to Jen, who has whilst I've been writing this thesis made me feel like I should be working a bit harder.

List of Contents

List of Figures

List of Tables

List of Equations

Abbreviations

%	Percentage
°C	Degrees centigrade
μ	Micro
Å	Angstrom
ADH	Alcohol Dehydrogenase
AKR	Aldo-Keto Reductase
APS	Ammonium persulfate
atm	Atmospheres
C	Centi
Da	Daltons
DMSO	Dimethyl sulfoxide
E_{act}	Active enzyme
E_{inact}	Inactive enzyme
eV	Electron volts
EXAFS	Extended X-ray Absorption Fine Structure
FFQ	Fast Flow Q-sepharose
g	Gram
IPTG	Isopropyl β-D-1-thiogalactopyranoside
K	Kelvin
k	Kilo
K_{cat}	Catalytic constant
k_d	Dissociation constant
K_{eq}	Equilibrium constant
K_{inact}	Inactivation constant
K_m	Michaelis constant
L	Litre
LB	Liquid broth
m	Metre
m	Milli
M	Molar
MAD	Multiwavelength anomolous dispersion
MDR	Medium chain dehydrogenase/reductase
MIR	Multiple isomorphous replacement
mol	Mole
MR	Molecular replacement
n	Nano
NAD^+	Nicotinamide adenine dinucleotide
NADH	Nicotinamide adenine dinucleotide (reduced form)
$NADP^+$	Nicotinamide adenine dinucleotide phosphate
NADPH	Nicotinamide adenine dinucleotide phosphate (reduced form)

OD$_x$	Optical density at wavelength x
Pcp	Pyrrolidone carboxyl peptidase
PIXE	Proton induced X-ray emission
rpm	Revolutions per minute
S	Seconds
SDR	Short chain dehydrogenases /reductases
SDS-PAGE	Sodium-dodecyl sulfate polyacrylamide gel electrophoresis
TEMED	Tetramethylethylenediamine
T$_{eq}$	Equilibrium temperature
THBA	1,2,3,6-tetrahydrobenzaldehyde
Tm	Melting temperature
TmAKR	*Thermotoga martima* aldo keto reductase
U	Units of activity
v/v	Volume per volume
Vmax	Maximum reaction velcoity
W	Watts
w/v	Weight per volume
ΔG^*_{inact}	Activation energy of inactivation
ΔG_{stab}	Energy of stabilisation
$\Delta\Delta G_{desolvation}$	Change in free energy of desolvation

Organism Abbreviations

A. pernix	*Aeropyrum pernix*
B. amyloliquefaciens	*Bacillus amyloliquefaciens*
B. halodurans	*Bacillus halodurans*
B. subtilis	*Bacillus subtilis*
C. tenuis	*Candida tenuis*
D. lebanonensis	*Drosophila lebanonensis*
E. coli	*Escherichia coli*
Pyroc. furiosus	*Pyrococcus furiosus*
S. acidalcaldarius	*Sulfolobus acidalcaldarius*
S. cerevisiae	*Saccharomyces cerevisiae*
S. solfataricus	*Sulfolobus solfataricus*
Thermoc. litoralis	*Thermococcus litoralis*
Thermot. martima	*Thermotoga martima*
Thermu. thermophilus	*Thermus thermophilus*

1 Introduction

1.1 Biotransformation

Biotransformations are able to catalyse a chemical reaction to produce a specific product by the use of a whole cell or an isolated enzyme. The most well known biotransformations are those carried out in the brewing and baking industries, where yeasts are utilised to produce alcohol during the brewing process or for CO_2 production to make bread light and airy. These processes have been utilised for thousands of years without any detailed knowledge of their mode of action. In the modern era, with a greater understanding of enzyme function, biotransformations are able to rival and even surpass many chemical synthetic reactions.

1.1.1 Historical Perspective

The nature and the date of the first biotransformations are unknown, but it is clear that biotransformation was being exploited before the beginning of recorded history. The history of brewing probably dates back to around 6000BC. Cheese production however may date back even further with some estimates putting it as early as 8000BC. The discovery of this particular biotransformation was probably an accidental one; sheep stomachs were commonly used to carry foods and if used for milk the rennet lining the stomach would begin the cheese making process.

Another ancient biotransformation was during the tanning of leather, where the process of bating was carried out by pounding the hide with animal dung or soaking in a solution of animal brains. These processes would introduce a mixture of amylases, proteases and lipases to the hide which would hydrolyse casein, elastin, albumin and globulin-like proteins leaving behind the collagen and keratin which make up the final leather.

1.1.2 Modern Biotransformations

The first successful preparative biotransformation was carried out in the 1930s, where *Acetobacter suboxydans* was utilised to convert *D*-sorbitol into *L*-sorbose as part of the synthesis of vitamin C (Reichstein and Grossner, 1934). In the 1950s and 60s the field began to grow, particularly with steroid hydroxylation which cannot be carried out by chemical means (Holland, 2000). The 1970s saw a trend begin for using isolated enzymes rather than whole cells. The vast majority of biocatalysts utilised as isolated enzymes since the 1970s have been lipases, mainly due to their robustness

and lack of cofactor requirements. With the advent of genetic engineering, recombinant DNA technology and genome sequencing the range of biocatalysts available has expanded, whilst the cost to produce them has decreased. When an appropriate biocatalyst has been found, its properties may be modified by site directed mutagenesis or directed evolution to provide a catalyst with improved stability or varied substrate specificity. Biotransformations are now utilised primarily in the production of a large range of pharmaceuticals in addition to finding applications in the fine and speciality chemical, foods, cosmetics, polymers and agrochemical industries (Fig 1.1) (Zaks, 2001; Huisman and Gray, 2002; Straathof, 2002). The biotransformation industry continues to grow as the technologies are developed and its potential realised. In 2000 the total enzyme market was worth $1.5billion (Kirk *et al.*, 2002), up from $1billion in 1995. By 2010, 70% of all chiral drug intermediates are expected to be produced using enzymes (Rouchi, 2002).
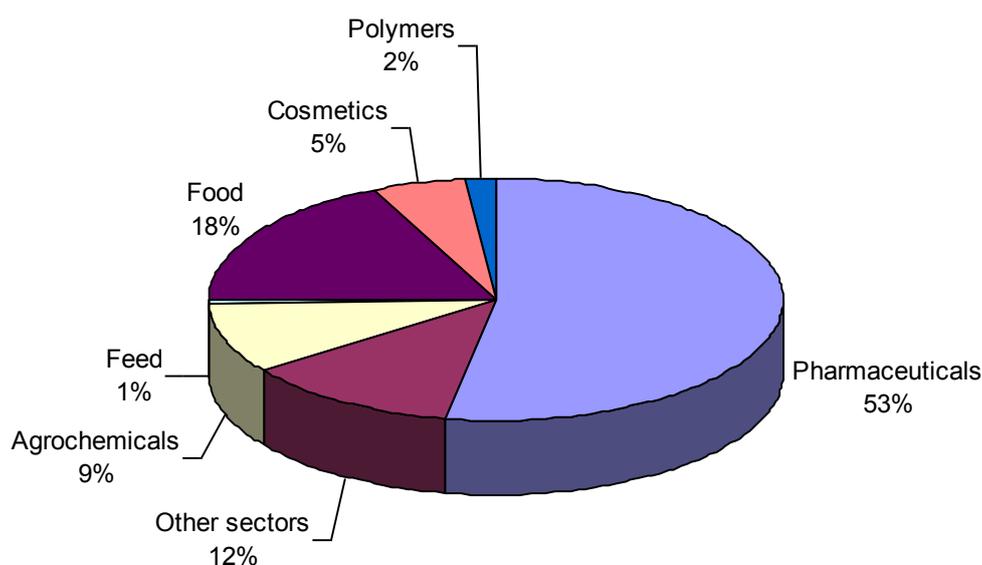
Fig 1.1. Application of biotransformation processes. Figure adapted from Straathof *et al.*, 2002.

1.1.3 Advantages and Disadvantages

1.1.3.1 Chirality

The production of chiral compounds by traditional chemical catalysis methods is either impossible or at best a lengthy and expensive process. This is where biotransformation has found its niche market. Due to the inherent chirality present in

biocatalysts many of their products are optically pure. Chirality exists when an object and its mirror image cannot be superimposed on each other. The simplest example of this is for a person's hands. The left and right hands are mirror images of each other and are non-superimposable, that is the left hand cannot be lined up in a way with the right hand so that all the features line up. There is an inherent difference between the two hands and they are both enantiomers of each other.

To aid in distinguishing between different enantiomers of a compound several naming conventions have been devised. These include the R/S convention based on the Cahn-Ingold-Prelog priority rules, where the substituent of lowest priority (lowest atomic number) is placed pointing away from the viewer and the order of priority of the remaining groups noted. If the order of priority (Fig 1.2a) is from highest to lowest in a clockwise direction it is labelled R (*Rectus*) or if it is anticlockwise then it is labelled S (*Sinister*). This system is commonly used in chemistry for its ability to distinguish between multiple chiral centres within an individual molecule. An alternative system is commonly used for amino acids and sugars, and uses glyceraldehyde as a reference. The two enantiomers of glyceraldehyde are labelled *D* and *L*. A simple way to distinguish between the two enantiomers of amino acids is by the CORN rule (Fig 1.2b), the chiral centre is viewed from the hydrogen and the positions of the **COOH**, **R** and **NH$_2$** groups observed (where R is the amino acid side chain). If the order proceeds in a clockwise direction then it is labelled the *L*-form or anticlockwise it is the *D*-form.

a)

b)
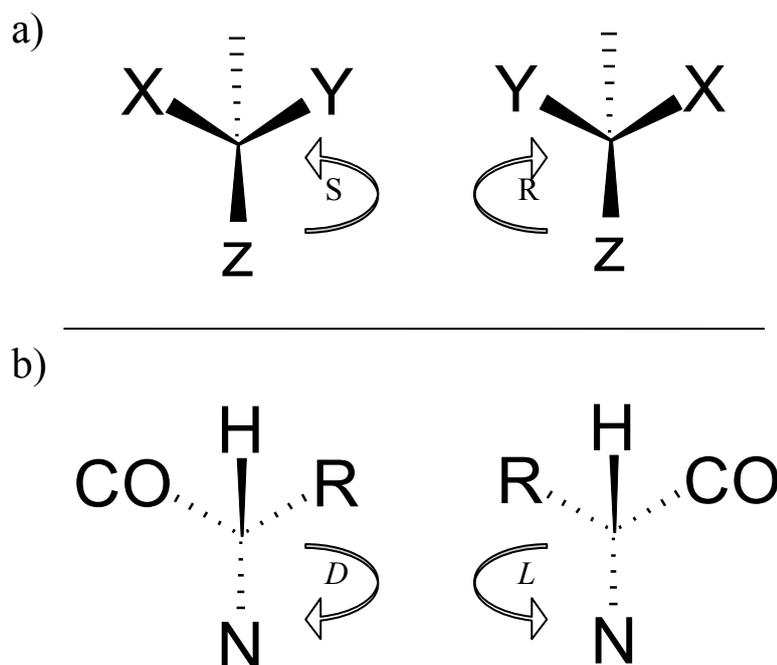
Fig 1.2. Naming conventions for chiral molecules. a) R/S convention, order of priority is z,y,x for highest to lowest. b) *D/L* convention using the CORN rule.

Chirality exists throughout the world of biology; most naturally occurring amino acids are of the *L* form and all sugars of the *D* form. *D* amino acids are found in bacterial cell walls. This preference for a single enantiomer leads to different interactions between a protein and the two enantiomers of a chiral compound. This can be demonstrated by the taste of amino acids, where the *D* amino acid tastes sweet the *L* form is tasteless. This is due to chirality present in the proteins which make up the taste receptors in a persons mouth distinguishing between the two enantiomers and responding only to the *D*-amino acids.

Since proteins are made up of amino acids of a single enantiomer (*L*), they have the ability to select between the two enantiomers of a substrate. This is termed stereospecificity. Stereoselectivity exists when the enzyme produces a single enantiomer of a product. Therefore a stereospecific process is always stereoselective but a stereoselective process is not necessarily stereospecific.

Enantiomerically pure products are especially important in the pharmaceutical industry, since only one enantiomer of a drug will usually have the desired effect and the other may have no effect or often cause side effects. This situation is exemplified by the case of thalidomide in the 1950s and 60s, where the S enantiomer is an anti-emetic effective against morning sickness and the R enantiomer is teratogenic causing

birth defects. In this case, use of the S enantiomer alone would not of prevented the side effects since the drug quickly racemises *in vivo* to give both the R and S enantiomers (Eriksson *et al*., 1995). Due to this case all drugs possessing chiral centres are now required to be produced in >99.9% enantiomeric excess or both enantiomers must be proven to be safe.

1.1.3.2 Other advantages

1.1.3.2.1 Regioselectivity

In addition to the chiral specificity of biocatalysts, they also possess regioselectivity. Regioselectivity is the ability to distinguish between the same chemical groups on a different part of a molecule. This allows an enzyme to distinguish between the multiple hydroxyl groups of a sugar molecule and act upon just one of them. To carry out a similar reaction using traditional chemical techniques would require multiple protection and deprotection steps to allow reaction with the desired functional group. The yield of the chemical method would decrease with each additional stage, whereas the biotransformation route could be carried out in a single step with the potential to reach 100% conversion. The regioselectivity of a biocatalyst is due to the 3-dimensional arrangement of the protein chain, and the particular shape of its active site. The shape of the active site will only allow the substrate to enter in a particular orientation thereby excluding the possibility of a reaction occurring on additional substituents. An additional advantage of the specificity provided by a biocatalyst is that fewer side products are formed reducing the need to remove these at a later stage and reducing the number of downstream processing steps required, further reducing waste and cost.

1.1.3.2.2 Green Chemistry

The conditions under which biocatalysts operate are usually very different from those of a chemical reaction. A chemical process will usually require high temperatures and pressures, extremes of pH, strong oxidising/reducing agents and are often performed in organic solvents. Biocatalysts are able to provide a 'green' alternative to chemical processes using relatively mild conditions; they are able to perform reactions with moderate temperatures and pressures, at mild pHs and in an aqueous environment. The reduction in temperature and pressure serves to reduce the operational costs of a process linked to heating the reactor and leads to a reduction in energy usage and

therefore reduced production of green house gasses. By removing the need for strong oxidising and reducing agents, extremes of pH and the use of organic solvents, there is a reduction in the risk to human health and the environment and a reduction in the cost of processing, clean-up and disposal of dangerous or harmful waste products. The use of a biocatalyst also removes the need to use toxic metal catalysts further reducing the environmental impact of a process.

As the use of biocatalysts for the production of chemical compounds increases so does the use of biocatalysts for removal of unwanted chemicals from the environment. Bioremediation of contaminants present in the environment further boosts the green credentials of biocatalysts, and this is a more cost effective and safer method of cleaning up environmental pollutants than by chemical means (Paul *et al*., 2005).

1.1.3.3 Disadvantages

Since biocatalysts are of biological origin and have generally evolved to work within the environment of the cell, they may behave unfavourably when introduced to process conditions. It is desirable for cost reasons to increase the volumetric productivity by increasing the substrate concentration. When the substrate concentration is increased a biocatalyst may loose activity, partially or wholly, due to inhibitory or toxic effects from the substrate or product. The cost of the biocatalyst may also be high compared to the cost of a chemical catalyst, due to isolation and stability of the biocatalyst. With improved cloning and expression systems the cost of producing a biocatalyst is reducing, and if immobilised this allows its easy recovery and costs are reduced even further. Immobilisation may also improve the operational stability of the biocatalyst by increasing its lifetime and profitability.

1.2 Oxidoreductases

Approximately one quarter of known enzymes are classified as oxidoreductases (Kula and Kragl, 2000). This large family of enzymes is structurally and catalytically diverse. They are capable of catalysing the oxidation of one substrate linked to the reduction of another substrate and are often called dehydrogenases. The most common oxidoreductases are those acting on alcohols and aldehydes. Oxidoreductases are of interest for use as biocatalysts for the production of chiraly pure alcohols, aldehydes and acids, production of polymers, and for use in biosensors and bioremediation (Hummel, 1999).

1.2.1 Cofactor Recycling

Enzymes such as oxidoreductases often require the use of a cofactor to carry out their reaction. These cofactors can often be more expensive than the desired product, and stoichiometric use of the cofactor would render the biotransformation process uneconomical. Due to the need for cofactor recycling there are few biotransformations carried out with purified oxidoreductases (Fig 1.3). An efficient method is therefore required to regenerate the cofactor to reduce the costs of the process and can also alter the thermodynamic equilibrium of the reaction driving it to completion (Findrik et $al.$, 2005). Of particular interest is the regeneration of nicotinamide adenine dinucleotide (NAD^+) and nicotinamide adenine dinucleotide phosphate ($NADP^+$) cofactors which are commonly utilised by the oxidoreductase enzymes. Cofactor regeneration can be achieved by a number of methods including chemical, electrochemical, photochemical, microbial and enzymatic. Probably the simplest method employed is the microbial approach which has been in use since brewing and baking were first introduced. This approach utilises the enzyme of interest within a metabolically active cell, and given an appropriate energy source the normal biochemistry of the cell is able to regenerate the cofactor. This method can often be limited by transfer of substrates and products across the cell membrane and may also suffer from side reactions due to the wide range of enzymes present with the cell. The enzymatic method of cofactor regeneration utilises a second substrate which is used to drive regeneration of the cofactor and is often the most convenient method of regeneration (Koeller and Wong, 2001). This step may be performed by the enzyme which carries out the synthetic step, and in which case it may be difficult to

achieve favourable reaction conditions for both the synthetic and regeneration reactions. Alternatively a second enzyme may be added to carry out the regeneration step (Fig 1.4). The second substrate may either be very cheap, easily regenerated or the second product be of value. The most common second enzymes and substrates utilised are formate dehydrogenase and formate, alcohol dehydrogenase and isopropanol and glucose dehydrogenase with glucose. The use of formate dehydrogenase with formate is attractive since the substrate if cheap, and the product is CO_2 which is easily removed as a gas and effectively makes the reaction irreversible, driving it to completion.
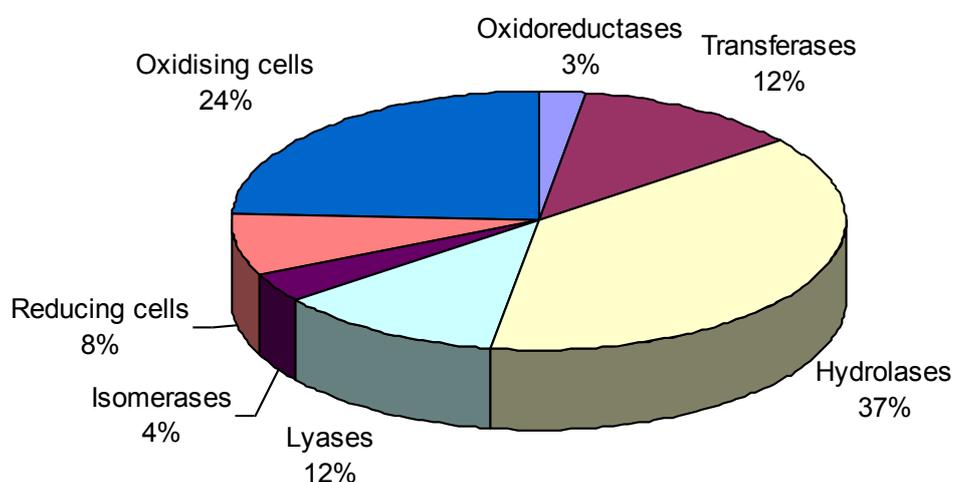
Fig 1.3. Types of enzymes used in industrial biotransformation reactions. Figure adapted from Straathof *et al*., 2002.
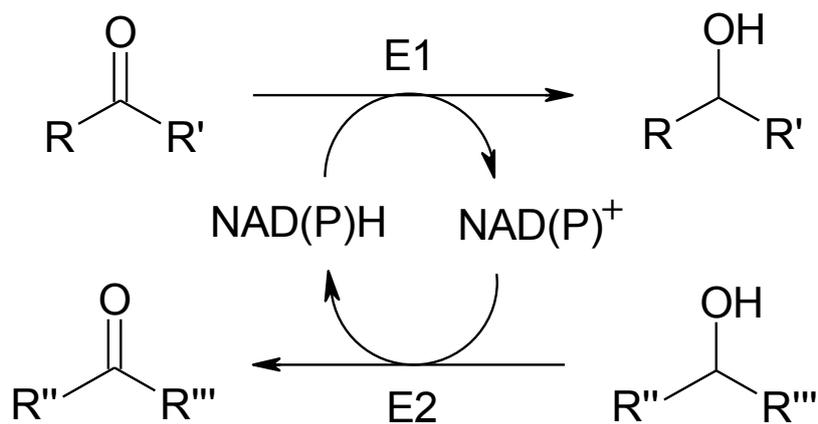
Fig 1.4. Enzymatic regeneration of NAD(P)H cofactor. E1 and E2 may be the same enzyme (substrate coupled regeneration) or different enzymes (enzyme coupled regeneration).

1.2.2 Alcohol Dehydrogenases

There are three evolutionary, catalytically and structurally distinct types of $NAD(P)^+$ dependant alcohol dehydrogenases (ADH). These are the classical alcohol dehydrogenases of which horse liver ADH is a member which are now referred to as medium chain dehydrogenases /reductases (MDR). There are also the short chain dehydrogenases /reductases (SDR) which are often called 'insect-like' since the first member of the family to be characterised was the *Drosophila lebanonensis* ADH (Villaroya *et al*., 1989). The third type are the long chain dehydrogenases or iron dependant ADHs which were originally identified in *Zymomonas mobilis* (Conway *et al*., 1987) and yeast (Williamson and Paquain, 1987). The structural diversity of the ADHs is shown in Figure 1.5. The iron containing ADHs are generally not well characterised and represent a much smaller enzyme family than the MDR and SDR families. They generally contain approximately 380 amino acid residues consisting of two domains, one with similarity to a dehydroquinate synthase like fold and the other a Rossmann fold like domain which binds the NAD(P)H cofactor. The catalytic iron is coordinated by three histidine residues and one aspartic acid residue.

The SDR family is a much larger protein family including lyases and epimerases in addition to the alcohol dehydrogenases. The enzymes within the SDR family are usually 270 amino acids in length and consist of a single Rossmann fold domain and they bind either NADH or NADPH and lack any metal cofactor. The active site contains a catalytic triad of Ser-Tyr-Lys, utilising the tyrosine as a catalytic acid-base.

The physiological function of the SDR family is though to be the reduction of aldehydes and ketones to their corresponding alcohols.

The MDR family contains the majority of the ADH enzymes and is characterised by the binding of one or two zinc ions and a chain length of 370 amino acid residues forming a two domain structure, with one substrate binding domain and one nucleotide binding Rossmann fold domain. The active site binds a catalytic Zn^{2+} and utilises serine and histidine in a proton relay system (Eklund *et al*., 1982), and a second zinc ion which has a structural role. Under physiological conditions this family usually acts in the oxidation direction, converting the alcohol to the corresponding aldehyde or ketone.
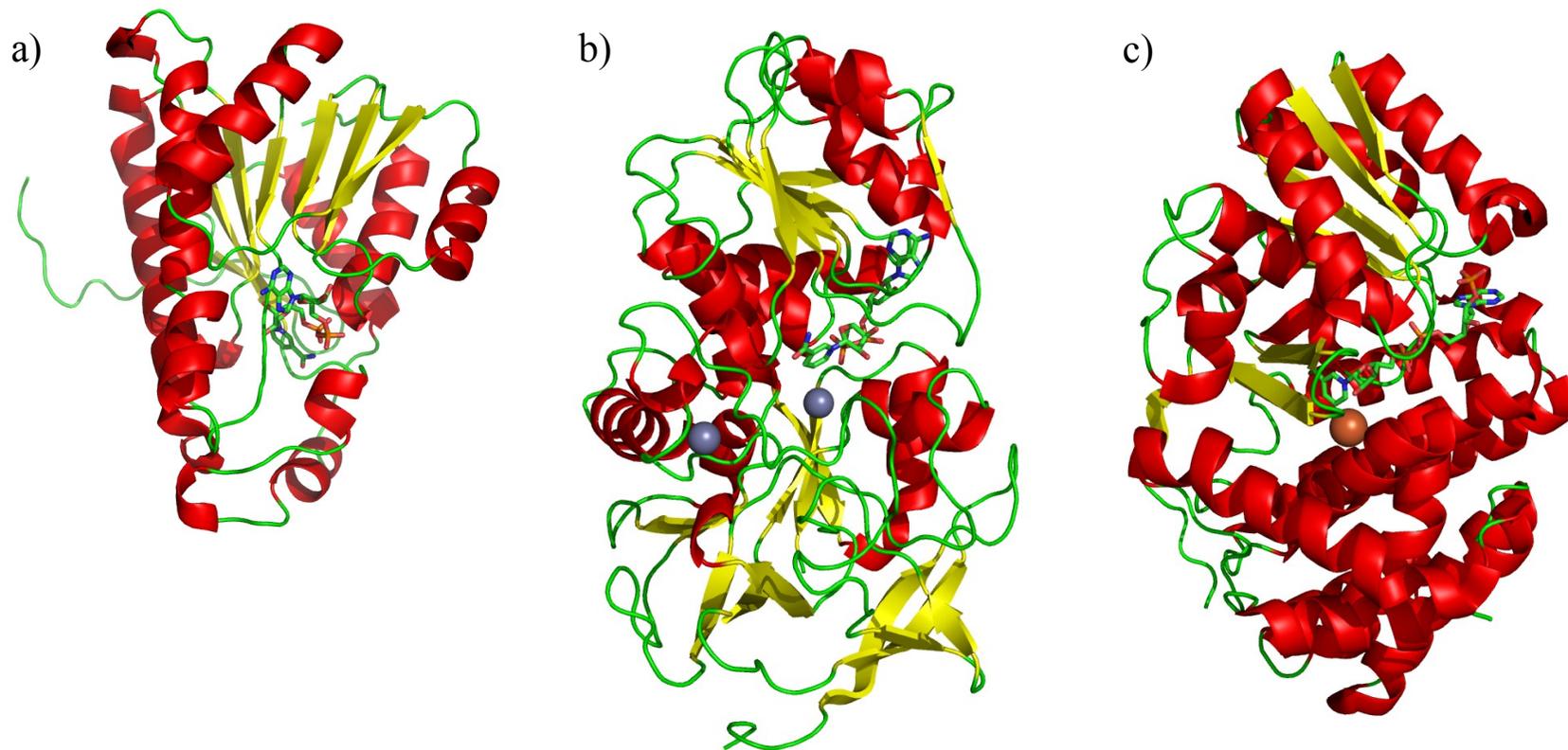
Fig 1.5. Figure showing the structural diversity of the alcohol dehydrogenase family. a) Short-chain dehydrogenase/ reductase family, SDR (*D. lebanonensis* alcohol dehydrogenase; PDB: 1B14). b) Medium-chain dehydrogenase/ reductase family, MDR (Horse liver alcohol dehydrogenase; PDB: 3BTO) c) Iron containing alcohol dehydrogenase family (*Thermotoga maritima* iron containing alcohol dehydrogenase; PDB: 1VLJ). The α-helices are coloured in red, the β-sheets in yellow and the loops in green. The zinc ions are shown as grey spheres, iron ions as red spheres and NAD(P)$^+$ as stick models. (Figure produced using PyMOL, Delano Scientific USA).

1.2.3 Aldo-Keto Reductase

It might be expected that the diversity present within the ADH family would fully account for the cells ability to reduce ketones and aldehydes. The aldo-keto reductase (AKR) family of enzymes is also capable of carrying out the same function and is present in all domains of life. The AKR family contains enzymes capable of reducing aldehydes and ketones as well as aldose reductases, xylose reductases and hydroxysteroid dehydrogenases. AKRs in mammals have been implicated as a factor in diabetic retinopathy, and are also though to play a role in the stress response. Many organisms contain multiple AKRs with overlapping substrate specificities which has lead to difficulties in assigning cellular functions. This is demonstrated by the yeast *Saccharomyces cerevisiae* which contains six different AKR genes. Knock-out studies performed on any one of these genes failed to yield a phenotype presumably due to functional redundancy between the six AKRs, when three of these genes were knocked out then a phenotype sensitive to heat stress was observed (Chang *et al.*, 2003).

The AKR family of enzymes are able to perform similar reactions to the ADH family described above, but are structurally very distinct. They are approximately 320 residues long, forming a single domain structure which folds into an $(\alpha/\beta)_8$ barrel which usually binds NADPH with the nicotinamide ring pointing towards the centre of the barrel (Fig 1.6). They are similar to the SDR enzymes in being single domain enzymes lacking any metal ion binding ability, but differ from all the ADHs by lacking a Rossmann fold like domain. Further similarity exists in the catalytic mechanism of the AKRs and the SDR enzymes. Both utilise a catalytic acid-base mechanism with a tyrosine residue as the general acid.

a)

b)
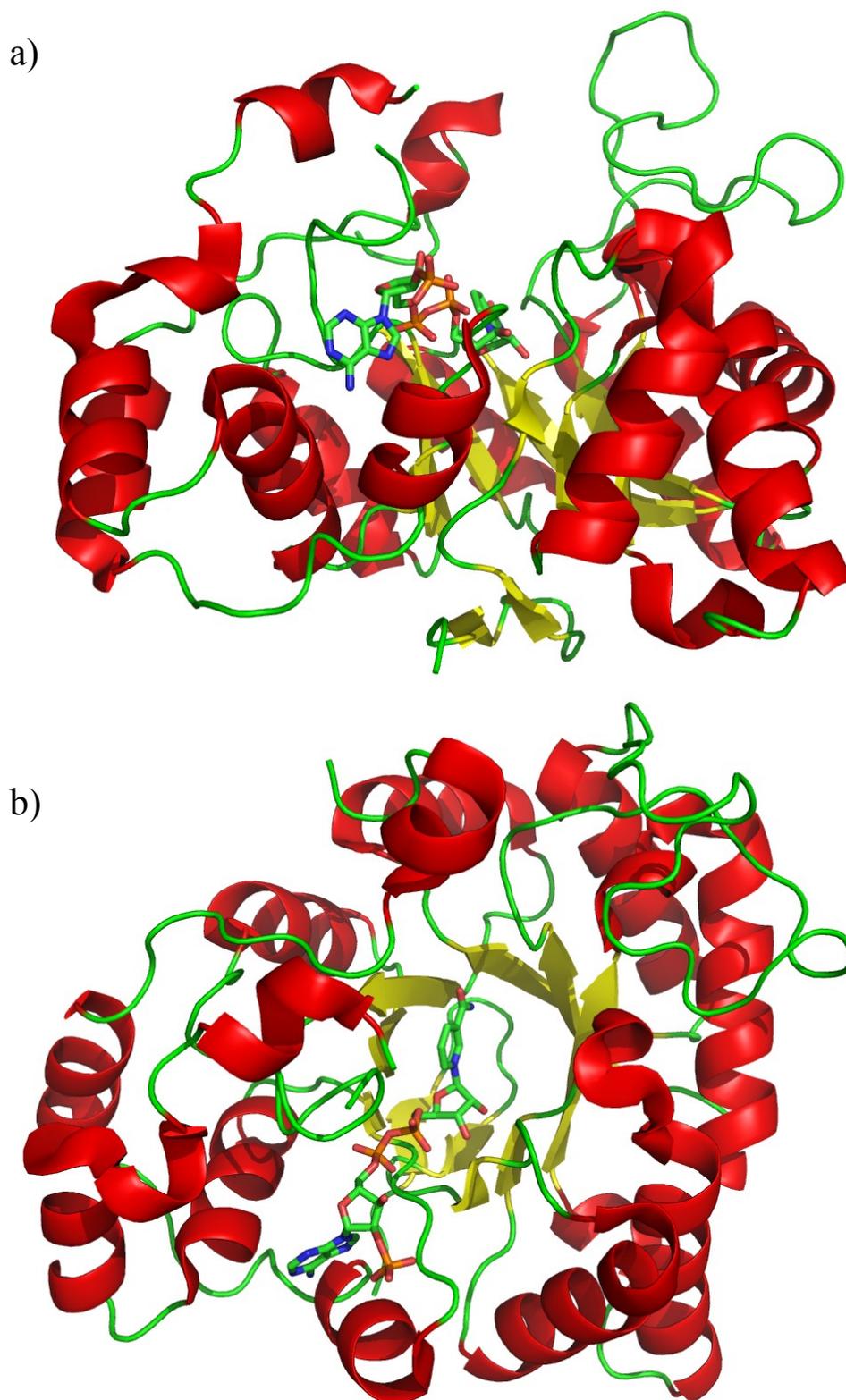
Fig 1.6. Figure showing the common structural fold adopted by the AKR family. The $(\alpha/\beta)_8$ barrel common to all AKRs with α-helices in red, β-sheets in yellow and NADP$^+$ as a stick model. Structures a) and b) are related by a 90° rotation around the x-axis. The structure shown is that of AKR2B5, *Candida tenuis* xylose reductase (PDB: 1K8C). (Figure produced using PyMOL, Delano Scientific USA).

1.2.3.1 Catalytic Mechanism

Within the active site of the AKRs is a catalytic tetrad (Fig 1.7) of residues (Tyr-Asp-Lys-His) in close proximity forming a hydrogen bonded network. Due to the proximity of the histidine and tyrosine to the C4 of the nicotinamide ring (Wilson *et al.*, 1992) it was not known which of these formed the catalytic residue. The tyrosine has now been established to be the catalytic residue by mutagenesis and kinetic studies on a range of family members (Mayr and Nidetzky, 2002; Penning, 1999; Schlegel *et al.*, 1998; Tarle *et al.*, 1993). The arrangement of the lysine and aspartic acid within the tetrad leads to a decrease in the pK_a for the tyrosine and facilitates proton donation to the substrate carbonyl. The location of the histidine may create an oxyanion hole (Jez *et al.*, 1997) which could stabilise an oxyanion intermediate. In all binary complexes of AKRs a water molecule is seen bound between the histidine and tyrosine residues at the location which would be occupied by the substrate carbonyl.
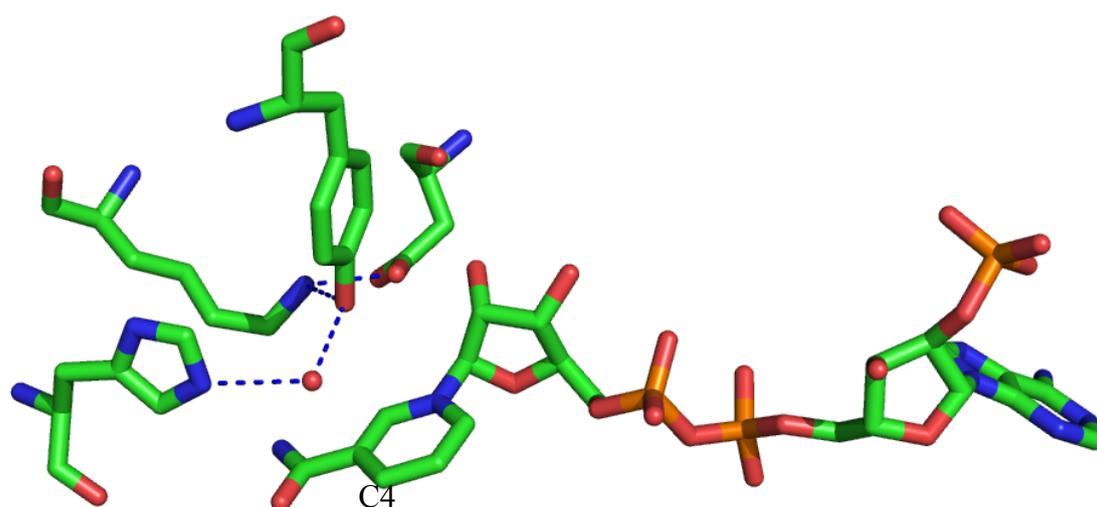
Fig 1.7. Figure showing the spatial arrangement of the catalytic residues and the cofactor within the AKR active site. A water molecule is found occupying the presumed location of the substrate carbonyl binding. Hydrogen bonds between the catalytic residues and water are shown as blue dashed lines. Structure shown is that of AKR2B5, *C. tenuis* xylose reductase (PDB: 1K8C). (Figure produced using PyMOL, Delano Scientific USA).

Cofactor and substrate binding occurs in a stepwise, ordered Bi-Bi method. In the rate limiting step, the NADPH cofactor binds first with the nicotinamide portion in the

centre of the barrel. Cofactor binding promotes a conformational change where the A and B-loops (loops 1 and 7 respectively, see chapter 6) close in around the cofactor and the B-loop specifically interacts with the pyrophosphate portion of the NADPH and locks it into position (Ehrensberger and Wilson 2004; Jez *et al*., 1997). The conformational changes observed upon cofactor binding have led to the loops A and B being termed the 'safety belt' due to the way they 'lock-in' the cofactor. After the cofactor has bound the substrate binds into the active site cleft created by the movement of loops A and B. After the reaction is complete, the product leaves first followed by the cofactor.

The spatial arrangement of the catalytic residues is highly conserved throughout the AKR family and the catalytic mechanism is therefore thought to be highly conserved too. It is thought that the catalytic mechanism proceeds partially or fully step wise with hydride transfer occurring before proton transfer (Nidetzky *et al*., 2001; Penning, 1999; Grimshaw *et al*., 1995). A hydride ion (H^-) is donated from the 4-*pro*-R position of the cofactor NAD(P)H to the substrate carbonyl forming an oxyanion intermediate stabilised by the nearby histidine residue (Fig 1.8). A proton (H^+) is then donated to the oxyanion intermediate from the catalytic tyrosine acting as a general acid to produce the alcohol product. The mechanism of the oxidation reaction proceeds in the reverse of the reduction reaction where the catalytic tyrosine acts as a general base (Mayr and Nidetzky, 2002). Some AKR family members show a large preference for reduction over oxidation (Nidetzky *et al*., 2001), making the reduction reaction almost irreversible.

The catalytic mechanism of the AKRs may be altered by a single amino acid change (Jez *et al*., 1997). The 3-oxo-5β-steroid 4-dehydrogenases contain a glutamic acid replacing the histidine of the catalytic tetrad; this substitution alters the activity of the enzyme from carbonyl reduction to carbon-carbon double-bond reduction. The substitution is thought to alter the mechanism by allowing the stabilisation of a carbonium-ion intermediate following proton transfer, with hydride transfer occurring by Markovnikov addition from the 4-*pro*-R position of NADPH.
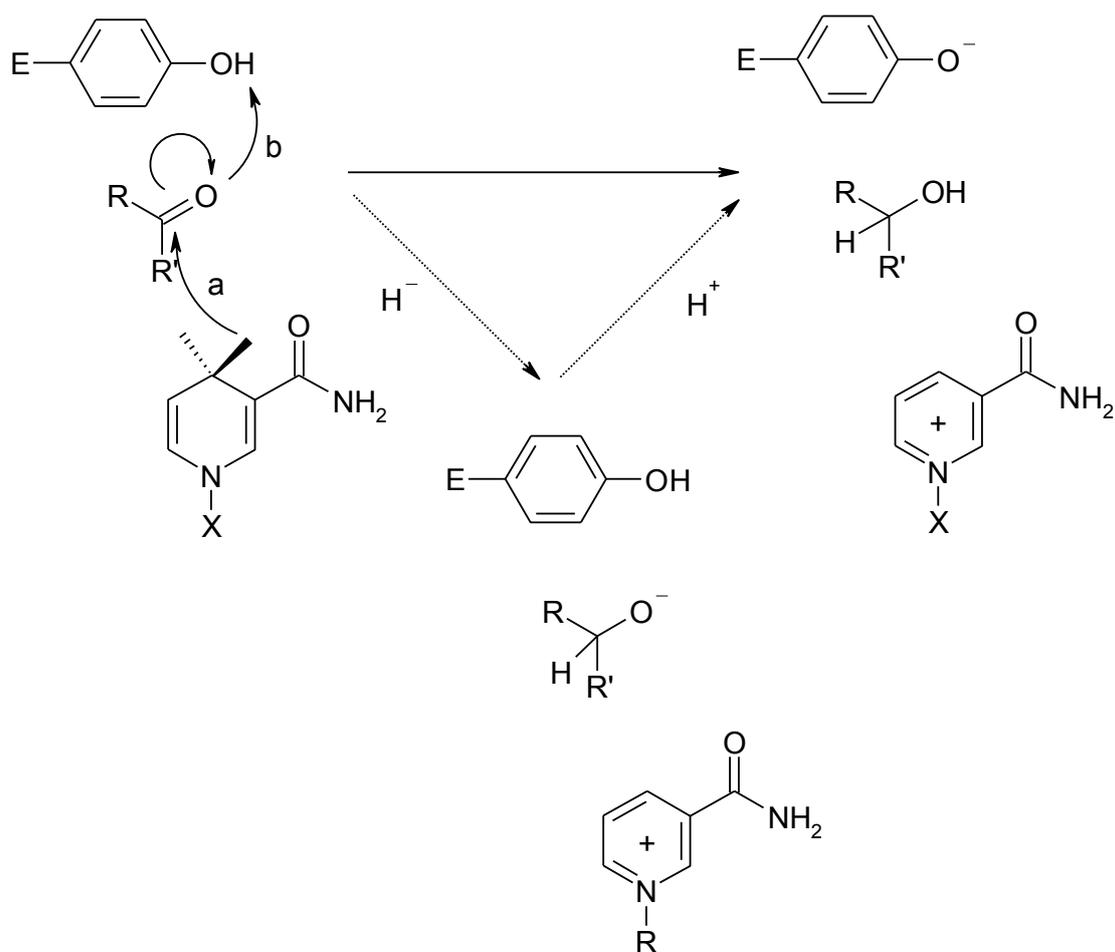
Fig 1.8. Figure showing the acid-base catalytic mechanism of aldehyde reduction by the AKR family enzymes. The reaction may occur fully stepwise with a) proton transfer from NADPH and formation of an oxyanion intermediate, prior to b) hydride transfer occurring from the enzyme tyrosine residue (Figure adapted from Mayr and Nidetzky, 2002).

1.2.3.2 AKRs in Biotransformations

The production of chirally pure pharmaceuticals is of high importance to the chemical industry, including the production of chiral secondary alcohols. There are several enzymatic routes available to achieve chirally pure alcohols including kinetic resolution by lipases, carbon-carbon bond formation by lyases and ketone reduction by oxidoreductases. The traditional ADHs have been exploited thus far, and have produced enzymes capable of yielding product with greater than 99% enantiomeric excess (Hildebrandt *et al.*, 2002). The AKRs represent an enzyme family which is capable of carrying out the reduction of a wide range of substrates, from steroids to sugars. The family all contain an $(\alpha/\beta)_8$ (TIM barrel) fold which is the most common

fold amongst known enzymes (Wierenga, 2001) making protein engineering decisions more reliable. Some of the enzymes within the family show preference for the reduction reaction to such an extent that the overall reaction is almost irreversible (Penning, 1999; Mayr and Nidetzky, 2002), driving any biosynthetic process towards completion.

In addition to chiral secondary alcohol synthesis for pharmaceutical manufacture the AKRs have roles in other industrially relevant processes such as vitamin C biosynthesis (Anderson *et al*., 1985; Ji and Goa, 2001), polyketide biosynthesis (Ikeda *et al*., 1999), xylose metabolism (Hahn-Hagerdal *et al*., 1991; Aristidou *et al*., 2000) and steroid metabolism (Takikawa *et al*., 1987). The AKRs are also of interest for their role in diabetic complications (Kador *et al*., 1988), cardiovascular disease and other related oxidative stress diseases (Uchida *et al*., 2000; Iuchi *et al*., 2004) where the human enzymes are of interest for inhibitor design.

1.3 Extremophiles

A mesophile is any organism which has optimum growth conditions which we consider 'normal', that is moderate temperature (20°C - 50°C), relatively neutral pH (3 - 9), low salt concentrations (<2M) and near atmospheric pressure. Any organism which thrives outside of these 'normal' conditions is considered an extremophile. The halophiles are able to survive in high concentrations of salt, and are found in places such as the Dead Sea and continental salt pans. Some extreme halophiles are able to survive in saturated salt conditions. Two methods are employed by the halophiles to survive such conditions. Either the salt can be excluded from the cell, in which case organic co-solutes are accumulated within the cell to increase the osmolarity and prevent efflux of water. Or alternatively the halophile may allow the salt to enter the cell and adapt their proteins to cope with such an environment (Madigan and Oren, 1999). Piezophiles are able to withstand high pressures, and are usually found on the ocean bed where pressures can reach over 1000atm (1000 times atmospheric pressure). At these high pressures the cell contents becomes more viscose so the organism must adapt all its components to become more flexible.

Acidophiles and alkaliphiles are able to resist extremes of pH, low or high respectively. The alkaliphiles survive by pumping protons from the environment across their membrane into the cytoplasm, allowing them to maintain a more neutral internal environment. The acidophiles may perform the opposite process, exporting protons to maintain a less acidic internal environment. Alternatively the acidophiles can adapt their enzymes to be resistant to such an environment. The thermophiles are organisms which grow between 50°C and 80°C; any organism able to grow beyond this temperature is classified as a hyperthermophile. The most extreme hyperthermophile dubbed 'strain 121' (Kashefi and Lovley, 2003) is able to grow within an autoclave at 121°C giving rise to its nickname. This organism has been found to remain viable up to 130°C, and was most closely related to the archaea *Pyrodictum occultum* and *Pyrobaculum aerophilum*. Thermophiles and hyperthermophiles are often found in geothermally heated regions of the world such as deep-sea volcanic vents, hot springs and solfataric fields. At the other end of the scale are the psychrophiles. These organisms grow optimally below 15°C and are often found in arctic regions. Generally the thermophiles have more interactions

increasing overall rigidity, whereas the psychrophiles have reduced rigidity and increased flexibility.

The majority of extremophiles are micro-organisms belonging to either the bacteria or the archaea. Several higher organisms also possess extremophilic properties, most notably the Pompeii worm (*Alvinella pompejana*) (Fig 1.9a) which lives at temperatures up to 80°C (Cary *et al*., 1998). The worm attaches itself by its tail to a hydrothermal vent and is thought to have a symbiotic relationship with the vent bacteria and archaea. In this way the tail is at 80°C and the head is at approximately 20°C. The Pompeii worm is able to live in temperatures up to 80°C, but an animal which usually lives under mesophilic conditions has proven to be the most extremo-tolerant organism known. The Tardigrades or 'water bears' (Fig1.9b) are usually found in moss and lichens but are able to survive almost any environmental extreme (Jonsson, 2007), including temperatures as low as -200°C for days or even -272°C (1K) for just a few minutes up to temperatures of 150°C, the low pressure of a vacuum up to 6000atm pressure (Seki and Toyoshima, 1998), high doses of radiation and desiccation for up to ten years. Tardigrades are able to survive such extreme environments by forming a 'tun' where the animal curls up into a ball, and replaces most of its water with trehalose and enters a state of crytobiosis (Goldstein and Blaxter, 2002) where its metabolism is 0.1% of normal.
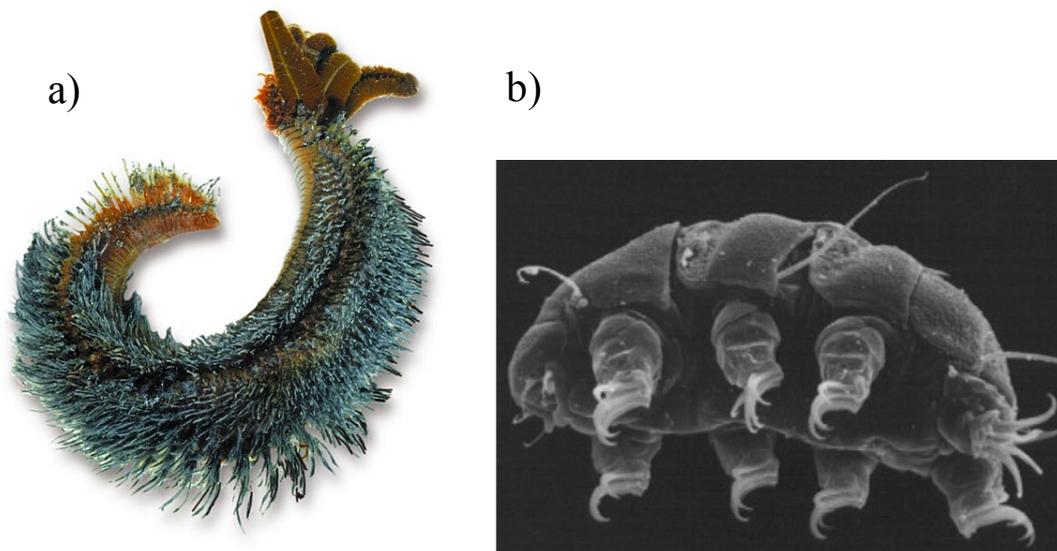
Fig 1.9. a) The *A. pompejana* or the 'Pompeii Worm' live with their head at 20°C and the tail attached to deep-sea hydrothermal vents at 80°C. (Image from http://www.udel.edu/PR/UDaily/2005/dec/extreme120304.html). b) SEM of a Tardigrade or 'water bear', usually found living in mosses. The water bear can survive almost any environmental extreme including space. (Image from http://www.fcps.edu/islandcreekes/ecology/water_bear.htm).

1.3.1 Thermophiles

The thermophiles and hyperthermophiles are often considered the most extreme microorganisms. As more become discovered the known limits within which life can exist is changing. The highest growth temperature found of any hyperthermophile is 121°C (see section 1.3) the previous record holder was *Pyrolobus fumarii* at 113°C (Blochl *et al*., 1997). The upper limit may never be conclusively known due to difficulties with culturing many microorganisms, but it is likely to lie between the known limit (121°C) and 150°C (Stetter, 1999). Life may find it hard to live beyond the current limit due to rapid degradation of many of the small biomolecules essential to life (Phipps *et al*., 1991) such as ATP and NAD^+ which hydrolyse rapidly at high temperatures. The temperature limit to life may depend upon the ability to prevent this degradation and ensure the rapid re-synthesis of these essential compounds.

Thermophiles have adapted themselves to living in an extreme environment through adaptation of their proteins (section 1.4) but also through their DNA and cell membrane lipids. The enzyme reverse gyrase is able to introduce positive supercoiling into the thermophiles DNA, which increases its stability (Kikuchi and

Asai, 1984). In addition the thermophilic archaea possess histone like proteins which can drastically increase the melting temperature of the DNA (Sandman *et al*., 1990). The cell membranes of the archaea are chemically unique compared to bacteria and eukaryotes (Fig 1.10). The lipids which make up the archaeal membrane are composed of a different enantiomer of glycerol (*L* rather than *D*), have ether linkages rather than ester linkages to the fatty acid tail and the tail is also chemically different being composed of isoprenoid units which create branching. The ether linkages in archaea are more stable to thermal and acid hydrolysis than the ester linkages of bacterial and eukaryal cell walls. The branching of the isoprenoid fatty acid tails may allow for two sidechains to join creating transmembrane phospholipids or intramolecular rings, both of which may increase the rigidity and stability of the cell wall and are often found in thermophilic species of archaea.
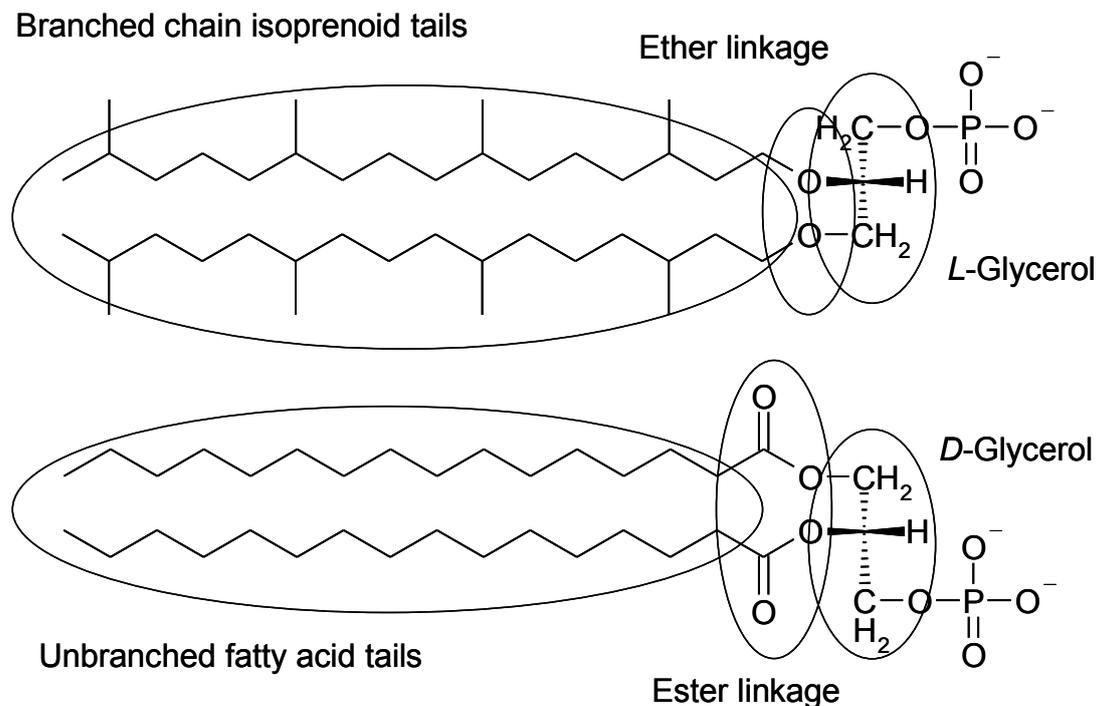
Fig 1.10. Figure showing the cell membrane phospholipids as found in a) archaea and b) bacteria and eukaryotes. The archaeal phospholipids contain *L* glycerol, with ether linked isoprenoid chains, whereas other organisms have *D* glycerol, with ester linked unbranched fatty acids.

1.3.2 Thermotoga maritima

Thermotoga maritima (Fig 1.11) is one of the most thermophilic species of bacteria, growing at a maximum of 90°C (optimum 80°C) metabolising simple and complex carbohydrates such as glucose, sucrose, starch, cellulose and xylan. Originally

isolated from geothermally heated marine sediment at Vulcano, Italy (Huber *et al.*, 1987). The organism is rod shaped and has a sheath like 'toga' structure (Fig 1.11) which forms an outer cell wall. The whole genome of *Thermot. maritima* has been sequenced (Nelson *et al.*, 1999) and is now the subject of a large scale structural genomics project (Canaves *et al.*, 2004). The complete genome sequence revealed a large proportion of the genome, 24%, showing high similarity to archaeal genomes. These archaeal like sequences are clustered into 15 regions within the genome, which maintain their gene order are probably a result of lateral gene transfer. Many of the archaeal like genes are associated with transport, electron transport, and conserved hypothetical proteins. Many of the genes associated with transcription, translation, DNA replication and cell division are related to the bacterial homologues. The *Thermot. maritima* also shares similarity with archaea in its cell membranes, where it contains novel glycerol ether lipids (see section 1.3.1) which increase the stability of the membrane at high temperatures compared to the ester lipid present in other bacteria.
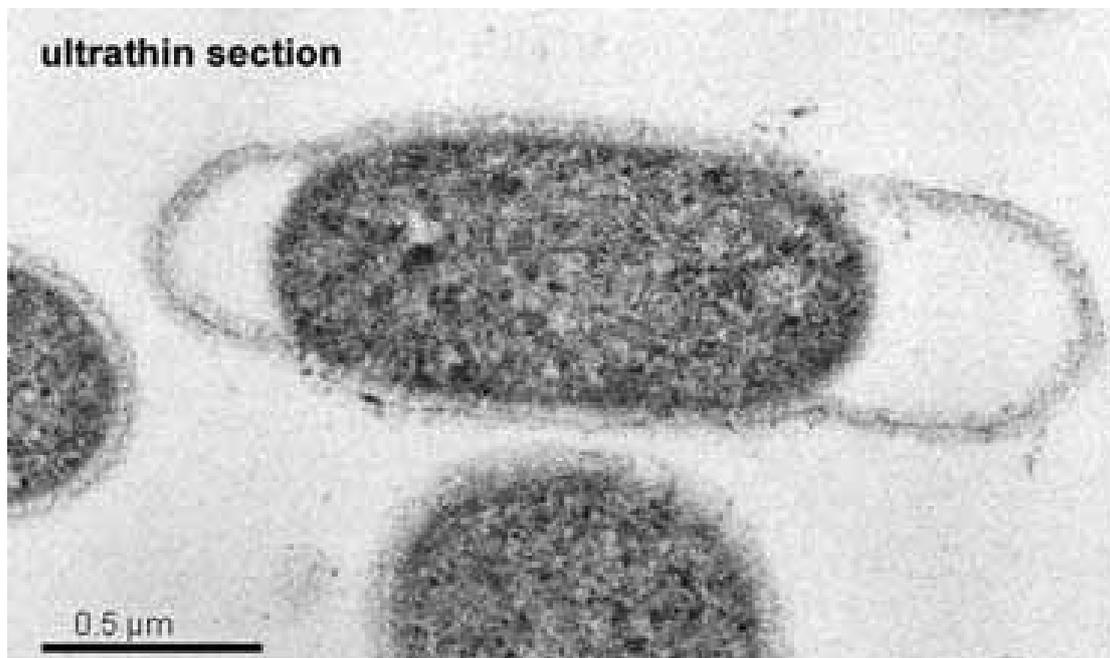
Fig 1.11. An ultrathin layer electron microscope image of a *Thermo. maritima* cell. Image taken from http://microbewiki.kenyon.edu/index.php/Petrotoga.

1.4 Protein Thermostability

An enzyme will retain its catalytic activity whilst it is able to maintain a conformation shape appropriate to its function. With the exception of disulfide bonds most proteins in the cytoplasm are held together by a combination of non-covalent forces including hydrogen bonds, ion pairs, hydrophobic interactions and Van der Waals interactions. As the temperature is increased these non-covalent forces fail to hold the protein structure together and the protein begins to unfold and denature. Once the protein is unfolded, hydrophobic residues which usually reside in the core of a protein may become exposed causing the protein to aggregate via interaction with other exposed hydrophobic residues. In addition, once unfolding occurs the protein is more susceptible to covalent modifications such as reduction of disulfide bonds, cysteine oxidation, deamidation of asparagine and glutamine residues and peptide bond hydrolysis.

Deamidation of asparagine and glutamine residues may occur by two different mechanisms (Wright, 1991), either a general acid-base mechanism or via a β-aspartyl mechanism. In the acid-base mechanism (Fig1.12a) the amide group of the side chain becomes protonated by a general acid. The carbonyl carbon of the protonated side chain is then attacked by a general base forming a tetrahedral oxyanion intermediate, which collapses to a carboxylic acid with removal of the amide group. In the β-aspartyl mechanism (Fig 1.12b) the peptide nitrogen residue n+1 acts as a nucleophile attacking the amide to produce a succinimide intermediate, which breaks down to form an α or β-linked amino acid. Peptide bond hydrolysis often occurs on the C-terminal side of asparagine residues by a β-aspartyl shift like mechanism (Fig 1.12c). The asparagine amide acts as a nucleophile, attacking the main chain carboxyl carbon of the same residue, breaking the peptide bond (Hensel *et al.*, 1992).

Under alkaline conditions β-elimination of the disulfide bond may occur, leading to production of thiocysteine and dehydroalanine (Fig 1.12d). The dehydroalanine may then further react with nucleophilic groups such as lysine to produce lysinoalanine. Reduction of disulfide bonds may occur under reducing conditions, breaking the bond and leading to further disruption of the protein structure. When not present as disulfide bonds, cysteines are the most reactive of all amino acids and are easily auto-oxidised and may lead to the formation of disulfide bonds or sulfenic, sulfinic or

sulfonic acid. Therefore the number of free cysteine residues is often reduced in thermophilic proteins, so this oxidation does not occur.

a) Deamidation via general acid-base mechanism
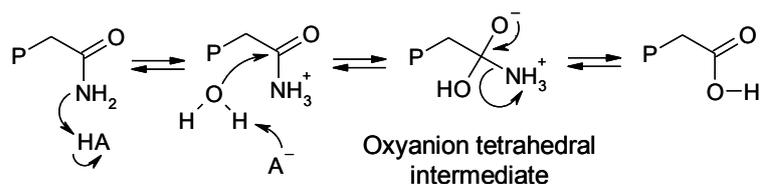
Oxyanion tetrahedral
intermediate

b) Deamidation via β-aspartyl mechanism

IsoAsp-Gly

Asn-Gly

Asp-Gly

c) Peptide chain cleavage

d) Disulfide bond cleavage
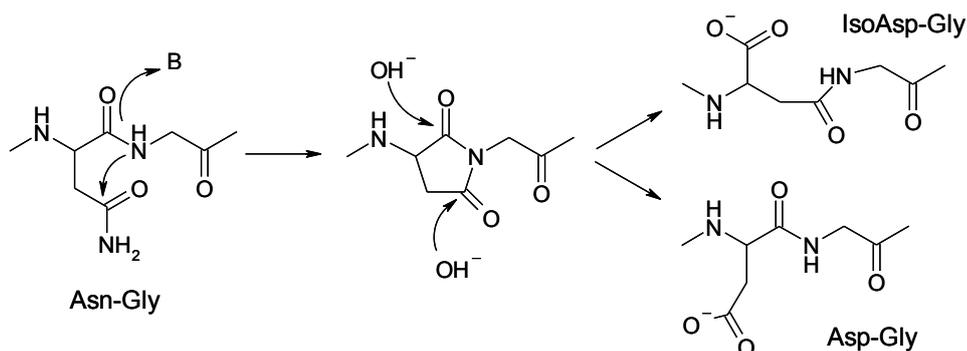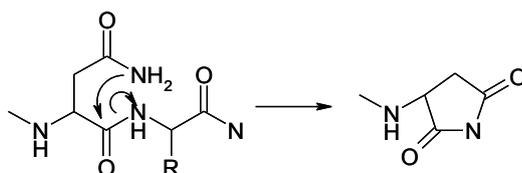
Disulphide bridge   Thiocysteine   Dehydroalanine
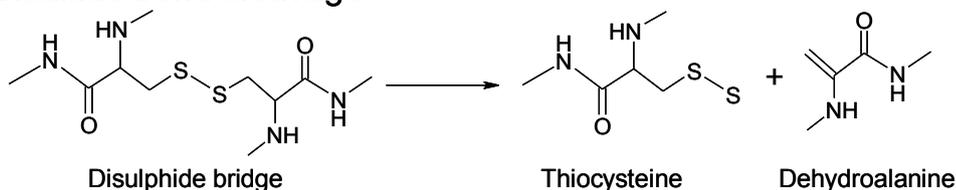
Fig 1.12. Figure showing the covalent modification mechanisms that occur in protein denaturation. a) Deamidation of asparagine and glutamine by general an acid-base mechanism, where the amide is protonated by a general acid, before a general base attacks the carbonyl carbon to form an oxyanion tetrahedral intermediate. b) Deamidation by the β-aspartyl mechanism where the side chain is attacked by a nucleophilic nitrogen of the neighbouring amino acid, forming a succinamide intermediate which breaks down to give either an α-linked (Asp) or a β-linked residue (IsoAsp). c) Peptide chain cleavage occurs by nucleophilic attack by side-chain amide onto its own peptide carbonyl leading to chain cleavage. d) Disulfide bond cleavage under alkaline conditions occurs via β-elimination and yields thiocysteine and dehydroalanine.

The unfolding and inactivation of a protein occurs in a way described by the equilibrium model of thermal inactivation (Eisenthal *et al*., 2006). The fully folded and active enzyme (E_{act}) is in a rapid equilibrium with a partially denatured inactive form (E_{inact}) which is on the pathway to fully denatured irreversible inactivation (X) (Eq 1.1). The kinetics of the pathway is described by the constants K_{eq} and K_{inact}. Where K_{inact} is the rate constant for irreversible inactivation of partially unfolded enzyme to fully denatured protein characterised by the activation energy of inactivation, ΔG^{*}_{inact}. K_{eq} is characterised by the enthalpy of equilibrium, ΔH_{eq}, and the thermal parameter, T_{eq}, which is the temperature at which the concentrations of active (E_{act}) and inactive (E_{inact}) enzyme are equal, thereby making T_{eq} the thermal equivalent K_m.

$$E_{act} \xrightleftharpoons{K_{eq}} E_{inact} \xrightarrow{K_{inact}} X \qquad \text{Eq 1.1}$$
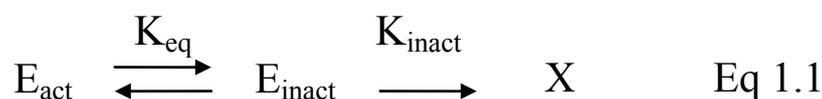
The overall stability of an enzyme may be affected by altering ΔG^{*}_{inact}, thereby altering the constant K_{inact}. The equilibrium model shows that an increase in ΔG^{*}_{inact} does not necessarily give an increase in activity at higher temperatures, since the equilibrium K_{eq} is not affected. An increase in T_{eq} however will shift the equilibrium in favour of E_{act} and increase the activity at elevated temperatures (Daniel *et al*., 2007). The energy difference between the folded state (both E_{act} and E_{inact}) and the fully denatured protein is represented by ΔG_{stab} and is usually equal to between 5-15kcal mol^{-1} for a mesophilic protein at 25°C (Jaenicke, 1991). ΔG_{stab} can be plotted against temperature to produce a profile for a given protein (Fig 1.13). In thermophilic proteins this ΔG_{stab} curve can theoretically be altered in three ways to give the thermophilic protein it's increased stability at higher temperature (Fig 1.13). Either the ΔG_{stab} value can be increased whilst still occurring at the same temperature optimum, the temperature of the highest ΔG_{stab} value can be shifted towards higher temperatures, or ΔG_{stab} may have the same value at the same optimum temperature but the curve may be flattened increasing stability at higher temperatures. Whilst mixtures of all three methods are commonly utilised by proteins to increase stability, the most commonly found mechanism is to increase the value of ΔG_{stab} (Vieille and Zeikus, 2001).
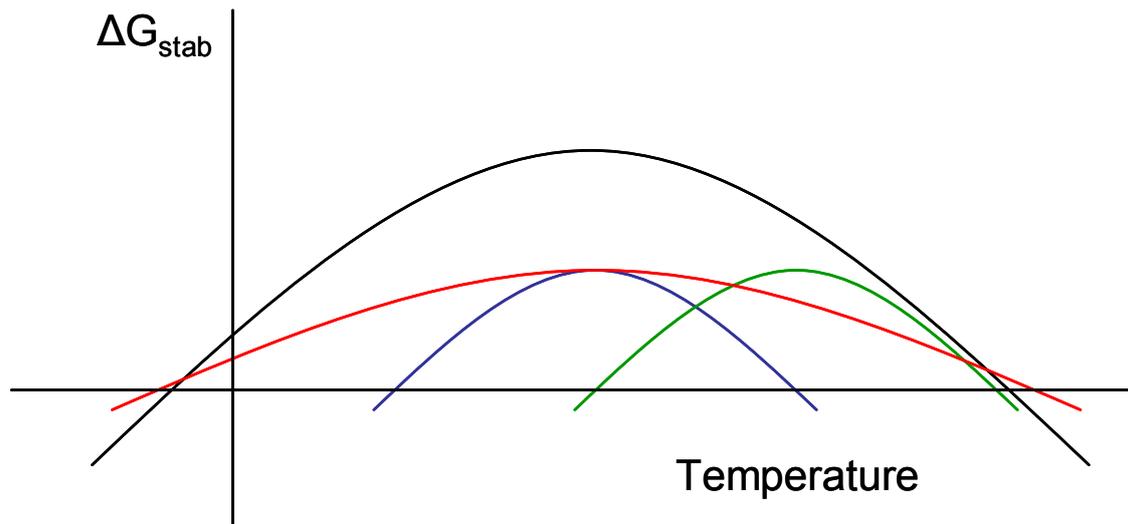
Fig 1.13. Figure to show the alteration of ΔG_{stab} vs temperature to give increased thermostability. The curve for a mesophilic protein is shown in blue. The three mechanisms for increasing thermostability are shifting the curve to higher temperatures (Green curve), increasing ΔG_{stab} value and keeping the same temperature optimum (Black curve) or peak ΔG_{stab} occurs at the same value but the curve is flattened (Red curve). Figure adapted from Vieille and Zeikus, 2001.

There are a number of structural adaptations which are known to increase the thermal stability of thermophilic and hyperthermophilic proteins. It has become clear from high resolution structures of thermostable proteins that there are no finite rules regarding which adaptations are utilised for thermostability, where one protein may utilise a particular mechanism of thermal stabilisation another may use an alternate mechanism to give the same degree of thermostability. The most commonly observed methods with which proteins achieve thermostability are discussed below.

1.4.1 Primary Structure

The relative abundance of certain amino acids has long been thought to affect the thermostability of proteins. Trends have been identified for increased incidence of charged residues, consistent with an increase in the number and size of ionic networks, and also for replacement of glycine with alanine and lysine with arginine. Alanine has a high propensity for forming α-helices, and would therefore increase stability by decreasing strain in α-helices. While arginine is thought to be less susceptible to chemical attack than lysine and is able mediate more ionic interactions increasing thermostability (Mrabet *et al*., 1992). Asparagine and glutamine, as

discussed previously (section 1.4), are readily deamidiated and might be expected to be of lower incidence in thermophilic proteins. Thermophilic phosphoglycerate kinase (PGK) enzymes have been shown to have a reduction in the number of asparagine residues in comparision to their mesophilic homologues (Hess *et al.*, 1995; Fleming and Littlechild, 1997). Also a trend for decreased asparagine and glutamine content with increasing optimum temperature was observed with type II xylose isomerases (Vieille *et al.*, 1995). Some thermophilic enzymes show the opposite trend, *Sulfolobus solfataricus* 5'-methylthioadenosine phosphorylase (optimally active at 120°C) contains twice as many asparagines as the *Escherichia coli* enzyme (Cacciapouti *et al.*, 1994) but probably protects them from deamidation by positioning them in locations and conformations which are not susceptible to deamidation.

Genome sequencing suggests however that overall there is not a preference for certain amino acids in thermophilic and hyperthermophilic genomes. In a comparison of eight mesophilic and seven hyperthermophilic genomes only minor trends were observed (Vieille and Zeikus, 2001), such as replacement of uncharged polar residues for charged residues, and an increase in hydrophobic and aromatic residues. Also a general reduction in cysteines was observed, especially in aerobic hyperthermophiles. The cysteines which were found in the aerobic hyperthermophiles were generally found to occur in essential interactions such as disulfide bridges, catalytic centres or metal chelation sites and were rarely found to occur otherwise.

1.4.2 Ion pairs

A study of 127 orthologous mesophilic/ thermophilic protein groups (Alsop *et al.*, 2003) found a clear link between thermostability and the presence of stabilising acid-base pairs on the surface of the protein. When an ion pair is formed there is an unfavourable change in the free energy of desolvation ($\Delta\Delta G_{desolvation}$), which may not be compensated for by the electrostatic energy of the ion pair. The $\Delta\Delta G_{desolvation}$ decreases as temperature increases, so the penalty is smaller for thermophilic proteins. In addition to forming large ion pair networks the desolvation penalty is further reduced since formation of each additional ion pair in a network requires the desolvation of only one additional residue. The ion pairs in thermophilic proteins are often seen in large networks of up to 24 residues (Yip *et al.*, 1995). The glucose dehydrogenase (GDH) from the thermophile *Pyrococcus furiosus* was compared to the GDH from the mesophile *Clostridium symbiosum*. The *P. furiosus* GDH has

fewer ionic networks containing less than 4 residues, but showed a large increase in the number of networks with more than 4 residues (Vieille and Zeikus, 2001). The ADH from *Aeropyrum pernix* is stabilised by ion pair networks at the subunit interfaces by four intersubunit ionic bonds (Fig 1.14) at each interface giving a total of 16 ionic bonds across the tetramer (Guy *et al*., 2003b). Another example of ionic networks in thermophilic proteins is the Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) from *S. solfataricus* which contains an ion pair network of 15 residues extending from the subunit interface to the active site (Isupov *et al*., 1999). This sequence is conserved in thermophilic GAPDH sequences but not in mesophilic homologues.
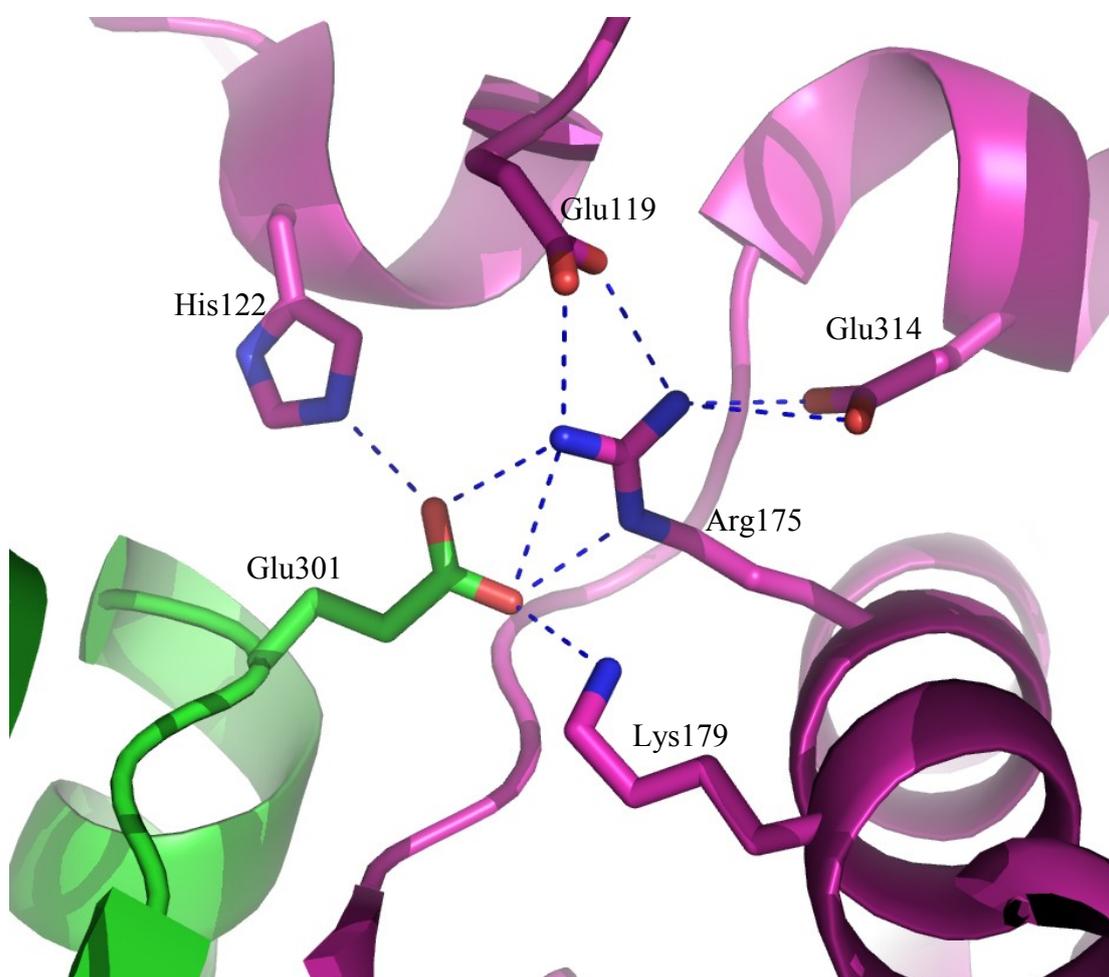
Fig 1.14. Figure showing the intersubunit ionic bonds in *A. pernix* ADH. Glu301 from one subunit interacts with His122, Arg175 and Lys179 from another subunit of the tetramer. Glu119 and Glu314 from the second subunit are also involved in the six residue ionic network. (PDB: 1H2B) (Figure produced using PyMOL, Delano Scientific USA).

1.4.3 Hydrogen Bonds

Hydrogen bonding in thermophilic proteins has not been extensively studied due to the lack of structures of sufficiently high resolution. The evidence available shows that hydrogen bonding can have a large contribution towards stability with charge-neutral hydrogen bonds thought to be the most important. A study on hydrogen bonding in GAPDH enzymes found a strong correlation between thermostability and the number of charge-neutral hydrogen bonds (Tanner *et al*., 1996). Such a hydrogen bond is formed between a side-chain atom of a charged residue and either a main-chain atom or a side-chain atom of a neutral residue. The hydrogen bond formed between two neutral residues is weaker than a charge-neutral hydrogen bond since the charge-dipole interaction is stronger than the dipole-dipole interaction. Charge-neutral hydrogen bonds have also been found to contribute to the stability of *Thermo. maritima* ferredoxin by stabilising the structure of turns and anchoring turns to one another (Macedo-Ribiero *et al*., 1996). The effect of hydrogen bonds on thermostability has been thoroughly studied in RNase T1, where 86 H-bonds were found and through the use of mutagenesis and calorimetry were found to contribute a similar stabilising effect to that of hydrophobic interactions (Shirley *et al*., 1992).

1.4.4 Disulfide Bridges

Disulfide bridges are able to stabilise a protein through an entropic effect, reducing the number of conformations the protein can adopt and hold different structural elements together through covalent modification. Disulfide bridges are however susceptible to hydrolysis at high temperatures, and lead many to believe that disulfides would not contribute towards stability in hyperthermophilic proteins. Disulfides have been found to be present in many proteins from hyperthermophiles with activity over 100°C. The 5'-methylthioadenosine phosphorylase from *S. solfataricus* when expressed in *E. coli* had a different arrangement of disulfide bridges and reduced activity suggesting that the disulfides in the native protein contribute towards the stability (Cacciapouti *et al*., 1994). Disulfide bridges in pyrrolidone carboxyl peptidase (Pcp) from *Thermococcus litoralis* were found (Singleton *et al*., 1999) which were not seen in the Pcp of the mesophile *Bacillus amyloliquefaciens* (Odagaki *et al*., 1999). When the cysteines were mutated to alanine the melting temperature and temperature optimum both dropped (Connelly, 2006). Conversely when these cysteines were introduced into the *B. amyloliquefaciens* Pcp the stability

was increased (Kabashima *et al*., 2001), suggesting a role for these cysteines in stabilising the protein.

1.4.5 Hydrophobic Interactions

Proteins bury their hydrophobic residues within their core and offer an increase in stability of 1.3 kcal mol^{-1} per methyl group buried away from the solvent (Pace, 1992). Evidence showing the importance of hydrophobic interactions in protein stability comes from a combination of X-ray structures and mutagenesis studies. Mutation of an isoleucine in the hydrophobic core of T4 lysozyme produced changes in the thermostability which correlated with the hydrophobicity of the residues substituted in place of the isoleucine (Matsumura *et al*., 1988). Hydrophobic interactions have been observed at subunit interfaces in the X-ray structures of many thermophilic proteins. The Pcp from *Thermoc. litoralis* utilises four hydrophobic residues from each subunit to stabilise its tetrameric structure (Singleton *et al*., 1999). Mutation of these residues has been shown to reduce the melting temperature of the enzyme (Littlechild *et al*., 2007). Also the 3-isopropylmalate dehydrogenase from *Thermus thermophilus* shows the addition of hydrophobic subunit interactions compared to the *E. coli* enzyme (Kirino *et al*., 1994). Mutation of these residues decreased the stability of the *Thermu. thermophilus* enzyme whilst introduction of hydrophobic residues at the same position in the *E. coli* enzyme increased stability. However the magnitude of the hydrophobic effect varies with temperature (Pace and Tanford, 1968). The effect of this is to reduce the hydrophobic contribution to stability at extremely high (and extremely low) temperatures. The Gibbs energy of hydrophobic interactions reaches a peak at 75°C (Makhatadze and Privalov, 1995), meaning that in hyperthermophiles the hydrophobic effect may be reduced at temperatures above 100°C. In the hyperthermophilic *Pyroc. furiosus* tryptophan synthase α-subunit, hydrophobic interactions in the protein interior did not contribute to the enzymes increased stability compared with the same enzyme from the mesophilic *Salmonella typhimurium*, instead the 33°C increase in melting temperature was attributed to an increase in ionic interactions (Yamagata *et al*., 2001). Some hyperthermophilic enzymes do still utilise increased hydrophobic interactions at temperatures around 100°C. A small 7kDa protein '7d' was studied from the hyperthermophiles *Sulfolobus acidocaldarius* (Sac7d; T_m = 90.7°C), and *S. solfataricus* (Sso7d; T_m = 99°C). The increased stability of Sso7d over Sac7d was

attributed to additional hydrophobic interactions within the core (Clark *et al*., 2004), demonstrating that even at such high temperatures hydrophobic interactions may still play a role.

1.4.6 Helix-Dipole Stabilisation

The α-helices in proteins contain an intrinsic dipole where the C-terminal end is partially negatively charged and the N-terminal end carries a partial positive charge. This dipole is induced by the directional nature of the helix as all the main-chain carbonyls are arranged pointing forward along the helix and the main-chain nitrogen faces backwards along the helix. By placing residues with a side-chain carrying an opposite charge to the dipole at the appropriate end of a helix, the capping residue may form hydrogen bonds to the main-chain of the helix stabilising the dipole (Fig 1.15). Dipole stabilisation has been observed in PGK enzymes from *Bacillus stearothermophilus* and *Thermot. maritima* in contrast to the yeast and pig enzymes (Auerbach *et al*., 1997). Both the thermophilic enzymes contained more helices stabilised at their N and C termini by oppositely charged residues. Helix dipole stabilisation may also be achieved by placing helices next to each other in an anti-parallel configuration so that their N and C termini pair up together and cancel out each others dipole. This has been shown to contribute to the stabilisation of a protein in contrast to packing helices in a parallel configuration which is destabilising (Sheridan *et al*., 1982). Helix dipole stabilisation has been attributed to increased thermostability, but its contribution is often marginal.
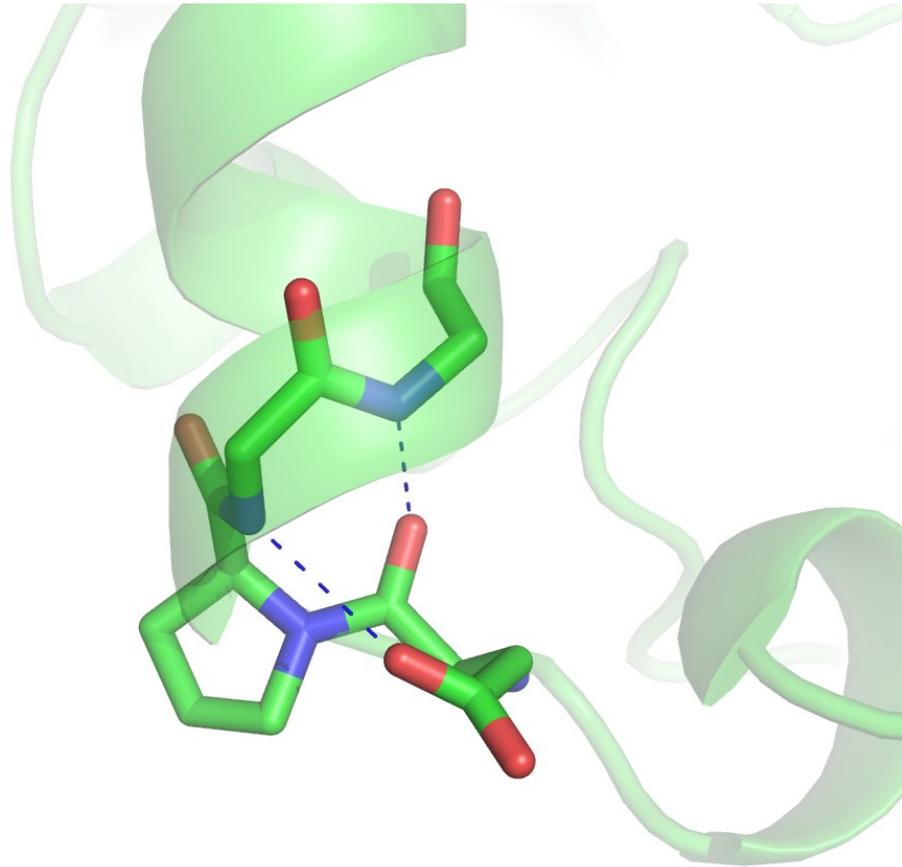
Fig 1.15. Figure showing the helix dipole stabilisation in *Thermot. maritima* phosphoglycerate kinase, showing the N-terminal cap with proline and aspartate stabilising the helix. The protein structure is shown as a cartoon representation with the residues involved in dipole stabilisation shown as sticks (PDB code: 1VPE) (Figure produced using PyMOL, Delano Scientific USA).

1.4.7 Structural Elements

There are a number of adaptations of the overall structure which are thought to contribute to thermostability. At a given temperature thermophilic enzymes are generally more rigid and less flexible than their mesophilic counterparts. Increased rigidity has been demonstrated experimentally through a reduction in hydrogen-deuterium exchange rates. At 20°C, less of a thermophilic enzyme's amide protons are exchanged for deuterium than a mesophilic enzyme, as the temperature is increased then the thermophilic enzyme begins to exchange protons (Bonisch *et al*., 1996). At each enzymes optimum temperature the percentage of exchanged protons is roughly equal, suggesting that at their optimum temperatures each enzyme has a similar amount of flexibility.

Loop structures are generally of reduced length in thermophilic proteins, while those that remain are often better anchored to the structure reducing conformational flexibility and the entropy of unfolding. The same applies to the N and C-termini, by anchoring the termini to each other or individually to the core conformational flexibility is reduced. The crystal structure of citrate synthase from *Pyroc. furiosus* forms a dimer and has an extension at the C-terminal end which wraps around the other monomer of the pair and is anchored in place (Russell *et al*., 1997). The extension wrapping around the other monomer increases the number of intersubunit interactions and is in contrast to the citrate synthases from less thermostable organisms. Better packing and reduction of internal cavities is not a common thermostabilisation mechanism but has been observed in some thermophilic proteins. This has been shown by site directed mutagenesis to contribute to the thermostability of histone proteins. In the mesophile *Methanobacterium formicium* the Ala31Ile and Lys35Met mutations increased the stability while the opposite mutations (Ile31Ala and Met35Lys) in the thermophile *Methanothermus fervidus* reduced its stability (Li *et al*., 1998). A reduction in the solvent accessible hydrophobic surface area has also been observed. In *S. acidocaldarius* superoxide dismutase, 18% of its accessible surface area is hydrophobic whereas in mesophilic enzymes 26.8% of the accessible surface area is hydrophobic (Knapp *et al*., 1999).

1.4.8 Extrinsic

The stability of a protein may be affected by the specific conditions under which it is studied. Intracellular enzymes reside within a controlled environment and may be stabilised by several extrinsic factors due to the composition of the cytosol. The presence of cofactor, substrate or some inorganic salts has been shown to increase the stability of a range of mesophilic and thermophilic enzymes. The effect of salt is dependant upon the enzyme and the salt present. Binding of cofactor or substrate is thought to act by protecting the active site from structural rearrangements. These factors appear to be as equally important to mesophilic proteins as they are in proteins from thermophiles. Pressure can also increase the thermostability of a protein (Robb and Clark, 1999). Since many thermophiles are naturally found in high pressure environments this may contribute to their stability. The stabilising effect is greatest in proteins stabilised by hydrophobic forces, and those stabilised by ionic interactions are often destabilised (Michels *et al*., 1997)

1.5 Protein Structure Determination

The 3-dimensional structure of biological macromolecules provides a wealth of data which can provide information on protein function, mechanism of action, interactions with other proteins, nucleic acids or small molecules and are able to provide information of distant evolutionary relationships not detectable by amino acid sequence similarity. The structures of proteins have also provided a large amount of information relating to mechanisms of thermostability (as discussed in section 1.4). If a structure is of high enough resolution and the protein participates in a disease mechanism then predictions can be made on the proteins interactions with drug molecules and structure based drug design carried out.

The first protein structure to be solved was that of sperm whale myoglobin (Kendrew *et al.*, 1958) which was solved using X-ray crystallography techniques. This first structure took nearly ten years to solve, but with developments in crystallisation techniques, X-ray sources, data collection techniques, and computer processor power the time required to solve a protein structure has been dramatically reduced. In 1971 the Protein Data Bank (PDB) was established with seven structures at the Brookhaven National Laboratory. On the 8th of April 2008 in its 37th year the PDB reached the milestone of containing over 50,000 structures (Fig 1.16). The growth of the PDB has been such that it has doubled in size since 2004, and with it currently receiving 25 new structures on average each day it is expected to triple in size by 2014 to contain over 150,000 macromolecular structures. Of the structures available in the PDB 85% were produced by X-ray crystallography methods (Fig 1.16), the other 15% were produced by NMR, electron microscopy and other methods (theoretical models are no longer stored in the PDB). The structures held within the PDB are freely available and are used by crystallographers, biochemists, molecular biologists and computational biologists from academia, industry and government.

Entries in PDB
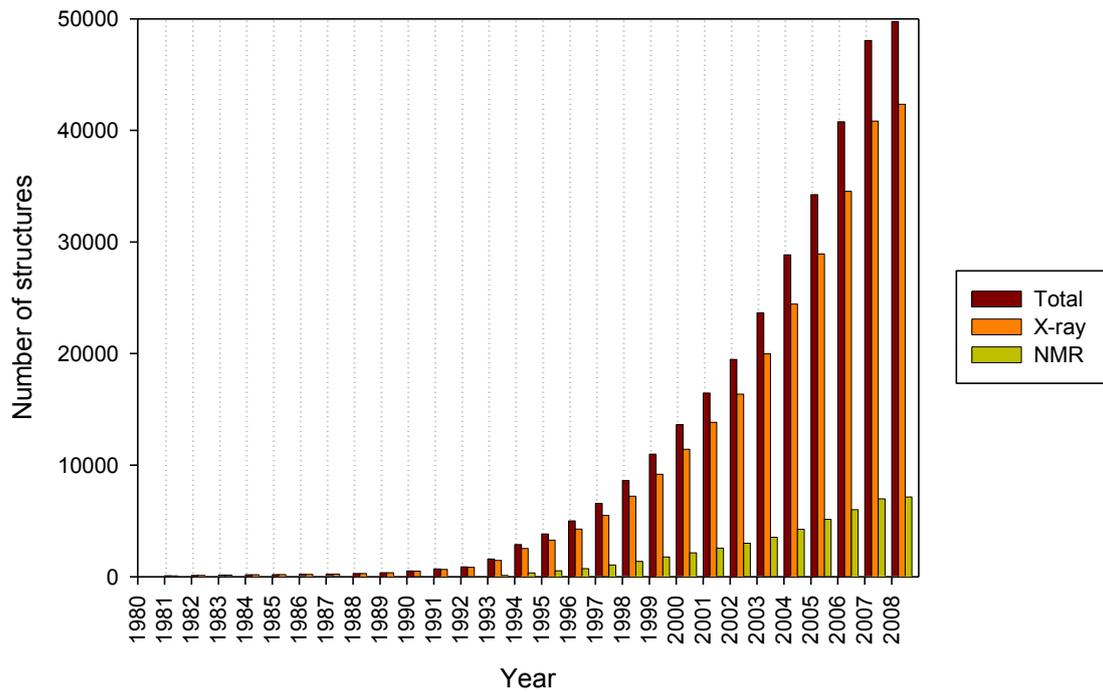
Fig 1.16. Figure showing the growth of the Protein Data Bank since 1980. The total number of structures in the PDB is shown in brown, with those solved by X-ray diffraction methods in orange and by NMR methods in mustard yellow. Figures for 2008 are year to 8th April. (Statistics from www.rcsb.org)

1.6 Previous Work

The work presented within this thesis is a continuation of a previous PhD project (Guy, 2004). This previous work included the cloning of an ADH from *A. pernix*, along with three putative oxidoreductases from *A. pernix*, *Thermot. maritima* and *P. furiosus*. The ADH belonged to the MDR family of zinc containing ADHs and was studied with respect to its biochemical and structural properties. It was found that the ADH was catalytically active towards a range of primary, secondary and cyclic alcohols and their respective ketones and aldehydes, showing a preference for cyclic alcohols. In addition to a broad substrate range the ADH enzyme is very stable with respect to temperature and organic solvents, showing a half life of 2 hours at 90°C and greater than 50% activity in 60% (v/v) acetonitrile or dioxane. X-ray crystallographic studies were carried out (Guy *et al*., 2003a) on the ADH and the structure solved by multi wavelength anomalous dispersion methods using the native zinc atoms (Guy *et al*., 2003b). The cofactor NADH and inhibitor octanol were found in the active site. This thesis has investigated the putative oxidoreductase from *Thermot. maritima*.

1.7 Aims of the Project

The oxidoreductase gene from *Thermot. maritima* previously cloned (Guy, 2004) has been investigated within this project. This oxidoreductase shows similarity to the Aldo-Keto Reductase (AKR) family of enzymes and is hereafter referred to as TmAKR.

The aims of the project were:

- To carry-out the expression and purification of the active TmAKR protein.
- To characterise the catalytic activity of the enzyme. Properties relating to the potential use of the enzyme will be particularly focused upon such as the substrate specificity and stereoselectivity and the stability of the enzyme.
- To obtain diffraction quality crystals of the enzyme.
- To solve the 3-dimensional structure of the enzyme by X-ray diffraction methods

The work presented within this thesis is the expression and purification, biochemical characterisation and X-ray crystallographic studies of the TmAKR enzyme. In addition the structure of *E. coli* bacterioferritin has been solved as a result of this project and is discussed within.

2 Expression and Purification

2.1 Introduction

To obtain large enough quantities of a protein for crystallisation a suitable source of the protein is required along with an efficient purification protocol to separate the protein of interest from any contaminant proteins. Before the advent of cloning and recombinant protein expression a protein could only be studied if it was present in high concentrations within the host organism or organ. An expression system consists of a host organism (bacteria, yeast, insect cell or mammalian cell) containing a vector harbouring the gene of interest. The choice of host organism depends upon the properties of the protein of interest and the requirements of the final product regarding secretion, glycosylation, proteolytic cleavage and other post-translational modifications. The vector carries along with the gene of interest promoters and regulators which control the level and time of gene expression, and antibiotic resistance markers allowing for selection of hosts containing the vector and the gene of interest.

After sufficient quantities of the protein have been produced it is necessary to separate it from the contaminating proteins present in the expression host whilst retaining the protein in an active state. Depending on the application it may be necessary to compromise on the purity or the yield of the protein. For instance when the protein is to be used for crystallisation a high quality and purity of protein is required therefore the overall yield may be sacrificed to remove impurities. Alternatively if the protein were to be used in a biocatalytic process obtaining high purity protein can be a waste of time and money so purity will usually be sacrificed for retention of the activity of interest.

Whatever the final application for the protein of interest it is important to choose the correct purification steps to give the most efficient procedure to help maximise yield. If a procedure consisting of five steps has a 75% yield of active target protein at each step then at the end of the procedure less than 25% of the target protein will be remaining. To provide an efficient protocol it is often useful to exploit a different property of the protein at each step. Common properties to exploit can be ionic charge of the protein, hydrophobicity, size/oligomeric state, heat stability, solvent resistance and ligand specificity.

2.2 Materials and methods

2.2.1 Transformation of TmAKR plasmid

The *Thermot. maritima* AKR (TmAKR) gene was cloned into the vector pTRC99 by Dr. Jodie Guy (Guy, 2004) and the purified plasmid was stored at -80°C. The plasmid was transformed into chemically competent *E. coli* BL21 by incubation on ice of 50 µL of competent cells with 1 µL of plasmid for 10 minutes, followed by heat shock for 60 seconds at 42°C before returning to ice for a further 2 minutes. 200 µL of LB medium was added and the samples incubated at 37°C for 1 hour prior to spreading the sample onto LB agar plates (with 0.1 g L^{-1} ampicillin) and incubated at 37°C overnight. Colonies were picked from the transformation plates and used to inoculate 10mL of LB media (with 0.1 g L^{-1} ampicillin) and grown overnight. Cultures were divided into 1mL aliquots and centrifuged (13000rpm, 10 minutes). Supernatants were removed and the pellets resuspended to 20% glycerol and flash frozen in liquid nitrogen prior to storage at -80°C until required.

2.2.2 Induction Studies

10 mL of LB medium containing 0.1 g L^{-1} ampicillin was inoculated with one glycerol stock of transformed *E. coli* BL21 containing the pTRC99 expression vector harbouring the TmAKR gene (see section 2.2.1) to provide an overnight starter culture. This culture was incubated with shaking (300rpm, 32mm orbit) at 37°C overnight for approximately 16 hours. Four 200 mL conical flasks containing 100 mL of LB medium and 0.1 g L^{-1} ampicillin were inoculated from an overnight starter culture. The flasks were incubated in an orbital shaker (300rpm, 32mm orbit) at 37°C and induced with 1mM Isopropyl β-D-1-thiogalactopyranoside (IPTG) at a specific OD_{595}. 1 mL samples were taken at regular intervals post induction to determine the optimum optical density at induction and the optimum post-induction time before harvesting. One flask remained uninduced to provide a basal level of expression, whilst the other flasks were induced at ODs of 0.6, 1 and 1.5. Induction was continued overnight. Samples were centrifuged at 13,000rpm for 10 minutes and the pellets stored at -20°C until all samples were collected. Pellets were resuspended in 200 µL of Bugbuster (Invitrogen) to lyse the cells and extract the soluble protein. Samples were centrifuged (13,000rpm, 10 minutes) to remove the cell debris and the soluble protein was analysed by SDS-PAGE.

2.2.3 Growth of Microorganisms

Six sterile 2 L conical flasks containing 1 L of LB medium and 0.1 g L^{-1} ampicillin were inoculated from an overnight starter culture of *E. coli* BL21 containing the pTRC99 expression vector harbouring the TmAKR gene. The flasks were incubated at 37°C, shaking (300rpm, 32mm orbit) for 8 hours (until OD$_{595}$ ~1.5) prior to induction of the TmAKR gene by addition of 1 mL 1M IPTG to each flask followed by incubation for a further 16 hours at 37°C with shaking (300rpm, 32mm orbit). Cells were then harvested by centrifugation at 12,000g for 20 minutes (4°C) and the pellets stored at -20°C until required.

2.2.4 Sodium-Dodecyl Sulfate Polyacrylamide Gel Electrophoresis (SDS-PAGE)

2.2.4.1 Stock Solutions

Acrylamide solution: 30% (w/v) in water (37.5 acrylamide: 1 bisacrylamide)

Stacking gel buffer: 0.5M Tris-HCl pH 6.8

Separating gel buffer: 1.5M Tris-HCl pH 8.8

SDS Solution: 10% (w/v) SDS in water

Ammonium persulfate (APS) solution: 10% (w/v) APS in water

Running buffer: 0.05M Tris-HCl, 0.38M glycine, 0.1% (w/v) SDS, pH 8.8

Sample loading buffer: 125mM Tris-HCl pH 6.8, 2% (w/v) SDS, 20% (v/v) glycerol, 0.001% (w/v) bromophenol blue, 0.005% (v/v) β-mercaptoethanol

Microwave coomassie stain: 0.05% (w/v) Coomassie brilliant blue, 50% (v/v) methanol, 10% (v/v) acetic acid, in water

2.2.4.2 Preparation of Gels

A Mini-protean® II gel apparatus (Bio-Rad Laboratories, UK) was set up as per the manufacturers instructions.   The separating gel (12.5%) consisted of 4.2 mL acrylamide solution, 2.5 mL separating gel buffer, 1 mL SDS solution, and 2.3 mL deionised water.   10 µL Tetramethylethylenediamine (TEMED) and 100 µL APS solution were added to begin polymerisation of the acrylamide, and the separating gel solution poured between the gel plates up to a height of approximately 50mm and water saturated isobutanol applied on top. Once set the butanol/water was washed off the top of the separating gel. The stacking gel (6%) consisted of 2 mL acrylamide solution, 2.5 mL stacking gel buffer, 1 mL SDS solution, 4.5 mL deionised water. 10

μL TEMED and 100μL APS solution were added to begin polymerisation and the stacking gel solution was poured onto of the separating gel to the top of the plates and a 10 well square toothed comb inserted and the gel allowed to polymerise.

2.2.4.3 Preparation of Protein Samples

20 μL of protein loading buffer was mixed with 20 μL of protein sample and incubated at 110°C for 20 minutes to ensure complete denaturation of the protein.

2.2.4.4 Molecular Weight Standards

To estimate the subunit size and easy identification of proteins during the purification, molecular weight standards were used. Precision plus pre-stained standards (Bio-Rad) were used which consist of 10 distinct protein bands with the following molecular weights 250, 150, 100, 75, 50 (Reference band), 37, 25, 20, 15, 10 kDa.

2.2.4.5 Running Procedure

The gel was inserted into the electrophoresis tank as per the manufacturer's instructions. Buffer reservoirs were filled with running buffer. Each sample to be analysed was loaded into a well in the gel using a Hamilton syringe and 5 μL of molecular weight standards was loaded into a well. The gel was run at 200 Volts for 55 minutes.

2.2.4.6 Staining Procedure

The gel was removed from the apparatus and placed into a 1 L beaker containing 50 mL of microwave coomassie stain and microwaved on full power (750W) for 3 minutes. The stain was removed and 700 mL of deionised water added and heated on full power for 10 minutes. After heating the gel was rinsed in deionised water and placed in fresh deionised water overnight to finish destaining at room temperature.

2.2.5 Activity Assay

The enzyme was assayed spectrophotometrically by monitoring the consumption of NADPH at 340nm using a Shimadzu UV2100 spectrophotometer at 50°C in a pre-warmed quartz cuvette. The reaction buffer, 50mM imidazole pH 6 (adjusted at assay temperature, and degassed), was incubated at 50°C for 5 minutes prior to performing the assay. 2 μL 1,2,3,6-tetrahydrobenzaldehyde (THBA) and 20 μL NADPH (5mM stock) was added to 960 μL of the preheated assay buffer and 20 μL enzyme sample

added to start the reaction. Assays were carried out for 5 minutes and the change in absorbance at 340nm followed. Assays were performed in triplicate and initial rates were calculated from the resulting curve and specific activities calculated using the protein concentration estimated from A_{280} readings and the extinction coefficient.

2.2.6 Protein Purification

All steps were performed at 4°C unless stated. Chromatographic procedures were performed on an ÄKTApurifier™ (GE Healthcare).

2.2.6.1 ExPASy

To aid in purification some parameters such as subunit molecular weight, isoelectric point (pI) and extinction coefficient (280nm) are useful to know. The protein sequence was input into the Protparam server (Gastieger *et al*., 2005) from ExPASy tools (Gastieger *et al*., 2003). The server output is shown in appendix 1.

2.2.6.2 Preparation of Cell Free Extract

Frozen cell pellets (section 2.2.4) were resuspended to 10% (w/v) in 20mM Tris-HCl pH 7.5. The optimum temperature and time period for the heat lysis step was determined by performing trials on 1 mL samples of the resuspended pellets and incubating these at a range of temperatures for varying periods of time. The samples were centrifuged (13000rpm, 10 minutes) and analysed by SDS-PAGE. For all following purifications the cell suspension was subjected to a heat lysis step at 80°C for 20 minutes and cell debris removed by centrifugation (12000g, 20 minutes).

2.2.6.3 Hydrophobic Interaction and Ion Exchange Chromatography

A 60 mL Phenyl-Sepharose hydrophobic interaction chromatography column (XK24; GE Healthcare) and a 10 mL Fast Flow Q-sepharose (FFQ) ion exchange column (XK16; GE Healthcare) were equilibrated with 2 column volumes of 20mM Tris-HCl pH 7.5, 2M KCl and 20mM Tris-HCl pH 7.5 respectively. The cell free extract was made to 2M with respect to KCl and centrifuged (12000g, 20minutes) to remove any particulate material. The cell free extract was then loaded onto the hydrophobic interaction column and washed with high salt buffer (20mM Tris-HCl pH 7.5, 2M KCl) to remove any unbound protein, and a gradient of 2-0M KCl over 400 mL was applied. At the end of the gradient the 10 mL FFQ was attached to the bottom of the first column and 180 mL of no salt buffer (20mM Tris-HCl pH 7.5) used to elute the protein from the Phenyl-Sepharose column and to bind it to the FFQ. At this point the

hydrophobic column was removed and a gradient of 0-0.5M NaCl applied over 10 column volumes of the FFQ to elute the protein of interest. 10 mL fractions of eluted protein were collected from the Phenyl-Sepharose column and 5 mL samples collected from FFQ and the A_{280} recorded throughout.

2.2.6.4 Size Exclusion Chromatography

Fractions containing the TmAKR protein were pooled together and concentrated to a total volume of 2 mL using a Vivaspin centrifugal concentrator (Sartorius, UK) with a 30kDa cut-off membrane and loaded onto a pre-equilibrated Superdex200 gel filtration column (XK16; GE Healthcare) with 20mM Tris-HCl pH 7.5, 0.1M NaCl. 120 mL of buffer was used to elute the protein and 1 ml fractions collected whilst monitoring the A_{280}.

2.2.6.4.1 Estimation of Native Molecular Weight

The elution profile from the Superdex 200 column can be used to predict the native molecular weight of a protein and therefore its oligomeric state. Seven proteins of known molecular weight are loaded onto the column and a calibration chart (Appendix 2) produced using LogMW and K_{av} values for each protein.

Kav is calculated as:

$$K_{av} = \frac{V_e - V_0}{V_t - V_0} \qquad \text{Eq. 2.1}$$

Where V_e = Elution volume, V_0 = Void volume, V_t = Total Volume

2.2.7 Bacterioferritin Purification

Frozen cells (section 2.2.4) were resuspended to 10% (w/v) in 20mM Tris-HCl pH 7.5. Resuspended cells were sonicated on ice using a Soniprep 150 (Sanyo) five times for 30 second periods with 30 second breaks to prevent overheating of the suspension. Sonicated cells were centrifuged (12000g, 20 minutes) to remove cell debris. The supernatant was incubated for 30minutes at 37°C with 1 U mL^{-1} benzonase and for a further 30 minutes at 4°C with the addition of 1 mg mL^{-1} protamine sulfate. The sample was centrifuged (12,000g, 20 minutes) to remove precipitated nucleic acids. The sample was then heat treated at 80°C for 20 minutes prior to centrifugation (12000g, 20 minutes) to remove precipitated *E. coli* proteins. Hydrophobic interaction chromatography, ion exchange chromatography and size

exclusion chromatography were all carried out as described for the TmAKR protein (sections 2.2.6.3 and 2.2.6.4)

2.3 Results and Discussion

2.3.1 Induction Studies

To optimise protein expression levels the post-induction time and the optical density at induction were varied and the results analysed by SDS-PAGE (Fig 2.1). Without any induction of the TmAKR gene the SDS-PAGE analysis shows a low level of expression of the gene. Induction levels were shown to increase with longer expression times and with initial induction at higher optical densities. An overnight induction (16 hours) with induction at an optical density of approximately 1.5 was chosen for all subsequent expression of the TmAKR protein.
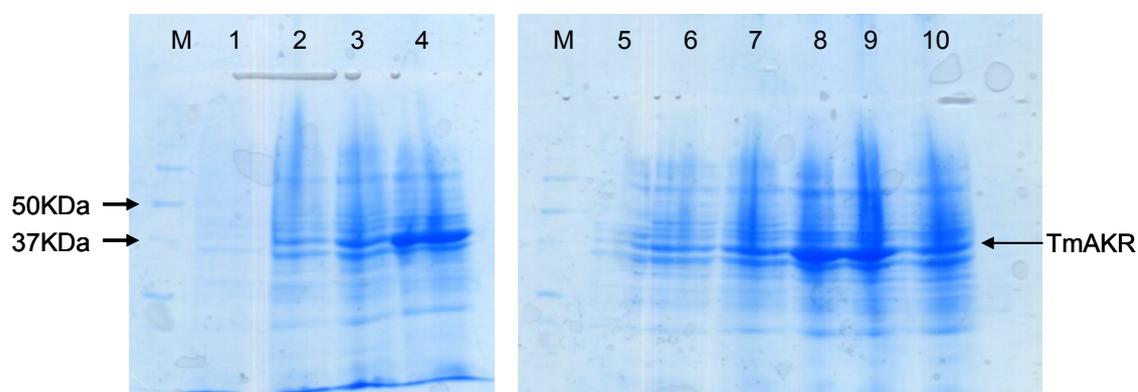
Fig 2.1. Figure showing the SDS-PAGE analysis of induction studies of TmAKR.

M – Protein molecular weight markers (see 2.2.5)
Lane 1 – Induction at OD 0.6, 0hrs induction
Lane 2 – Induction at OD 0.6, 2hrs induction
Lane 3 – Induction at OD 0.6, 4hrs induction
Lane 4 – Induction at OD 0.6, 20hrs induction
Lane 5 – Induction at OD 1.0, 0hrs induction
Lane 6 – Induction at OD 1.0, 2hrs induction
Lane 7 – Induction at OD 1.0, 4hrs induction
Lane 8 – Induction at OD 1.0, 19hrs induction
Lane 9 – Induction at OD 1.5, 16hrs induction
Lane 10 – No induction 24hrs growth

2.3.2 Purification

From the program Protparam, the subunit molecular weight of the TmAKR was calculated to be 37.5 kDa (37538 Da), the pI was estimated as 5.56 and the extinction coefficient at 280nm to be 41160 M^{-1} cm^{-1} (Absorbance 1 g L^{-1} = 1.09). Knowing the subunit molecular weight allows us to identify the protein throughout purification on SDS-PAGE more reliably. The extinction coefficient can be used to estimate the

protein concentration from the absorbance at 280nm. The predicted pI of 5.6 can be used to design the purification procedure. By using a buffer with a pH of approximately 7.5 the protein should be negatively charged and be able to bind to an anion exchange resin. To simplify the procedure and remove the need to exchange buffers, Tris-HCl with a pH of 7.5 was chosen to use throughout the purification.

The source organism for the AKR, *Thermot. maritima* has an optimum growth temperature of 80°C and therefore many of its enzymes would be expected to be stable up to and above this temperature. Due to this property of the enzyme it is possible to heat up the sample to high temperatures and remove large quantities of the native *E. coli* proteins. Heat lysis trials were performed at a range of temperatures and times and the remaining soluble protein analysed by SDS-PAGE (Fig 2.2). The SDS-PAGE analysis shows that at 80°C a large proportion of the AKR remains but at higher temperatures and longer times the AKR is slowly denatured without significantly removing extra *E. coli* proteins. Incubation at 80°C for 20 minutes provided the best compromise for removal of *E. coli* proteins and retention of the TmAKR activity. By performing this procedure on the intact *E. coli* cells it was possible to disrupt the cell wall and carry out the first two stages of the purification in one step.
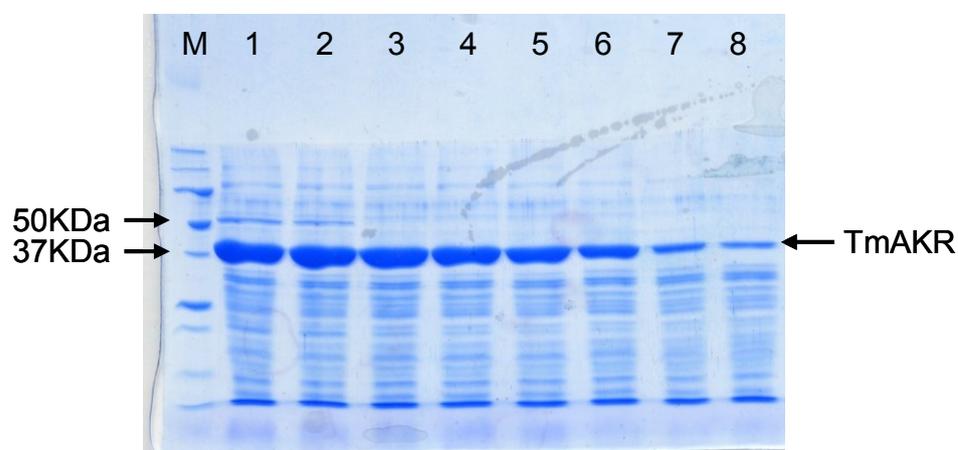
Fig 2.2. Figure showing the SDS-PAGE analysis of heat lysis trials of TmAKR.

M – Protein molecular weight markers (See 2.2.5)
Lane 1 – 80°C, 20 minutes
Lane 2 – 80°C, 30 minutes
Lane 3 – 85°C, 10 minutes
Lane 4 – 85°C, 20 minutes
Lane 5 – 85°C, 30 minutes
Lane 6 – 90°C, 10 minutes
Lane 7 – 90°C, 20 minutes
Lane 8 – 90°C, 30 minutes

The semi-purified protein from the heat lysis step was loaded onto a Phenyl-Sepharose column and the TmAKR bound successfully to the column as can be seen from the activity profile (Fig 2.3) and SDS-PAGE (Fig 2.4) analysis of the eluted proteins. There is a small amount of activity at the end of the gradient from the Phenyl-Sepharose column before the FFQ is connected; the SDS-PAGE shows that this activity is due to the presence of the TmAKR. A minimal amount of protein is lost at this stage which is not recovered; this loss is outweighed by the gains made by performing these two column chromatography steps in tandem. The TmAKR can be directly eluted onto the FFQ column since it elutes at the end of the salt gradient and contains very little salt to prevent it binding to the ion exchange media. The AKR elutes from the FFQ column approximately half way through the gradient in a roughly symmetrical peak (Fig 2.5) showing the success of the first two purification steps and is easily identified by its activity and presence on SDS-PAGE (Fig 2.5, 2.6).
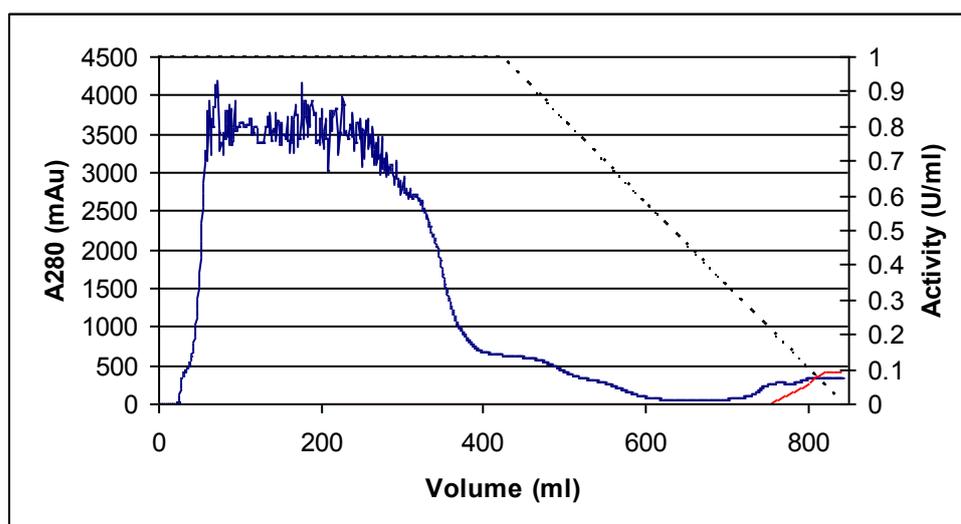
Fig 2.3. Figure showing the A_{280} elution profile (Blue) and Activity profile (Red) of hydrophobic interaction chromatography of the TmAKR protein. The KCl gradient is shown in a black dashed line from 2M – 0M (scale not shown)
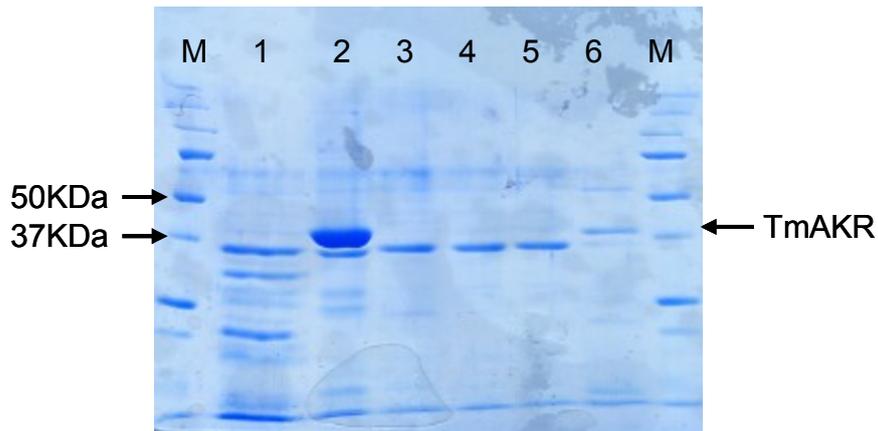
Fig 2.4. Figure showing the SDS-PAGE of hydrophobic interaction chromatography, and ion exchange chromatography of TmAKR.

M – Protein molecular weight markers (See 2.2.5.)
Lane 1 – HIC 200-210ml
Lane 2 – HIC 800-810ml
Lane 3 – FFQ 10-15ml
Lane 4 – FFQ 55-60ml
Lane 5 – FFQ 105-110ml
Lane 6 – FFQ 210-215ml
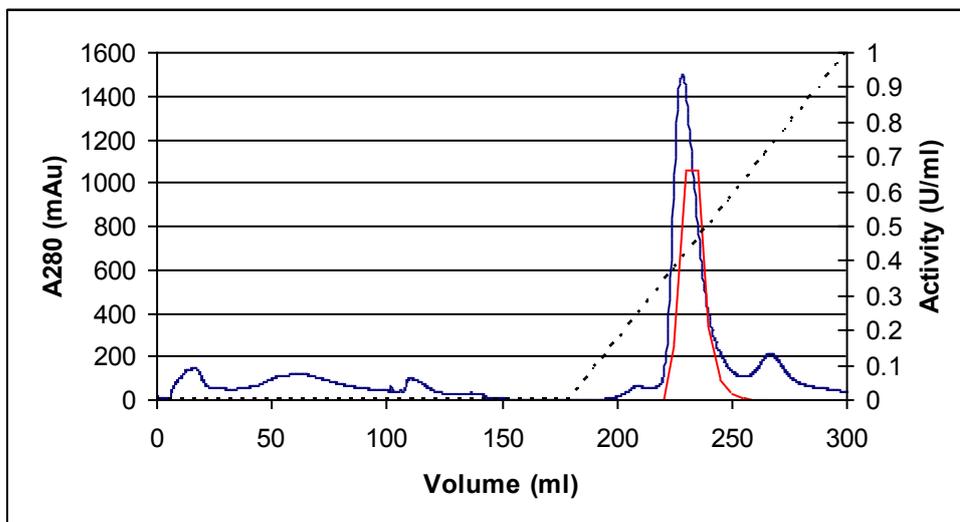
Fig 2.5. Figure showing the A_{280} elution profile (Blue) and Activity profile (Red) of ion exchange chromatography of the TmAKR protein. A gradient of 0 – 0.5M NaCl is shown in a black dashed line (scale not shown)
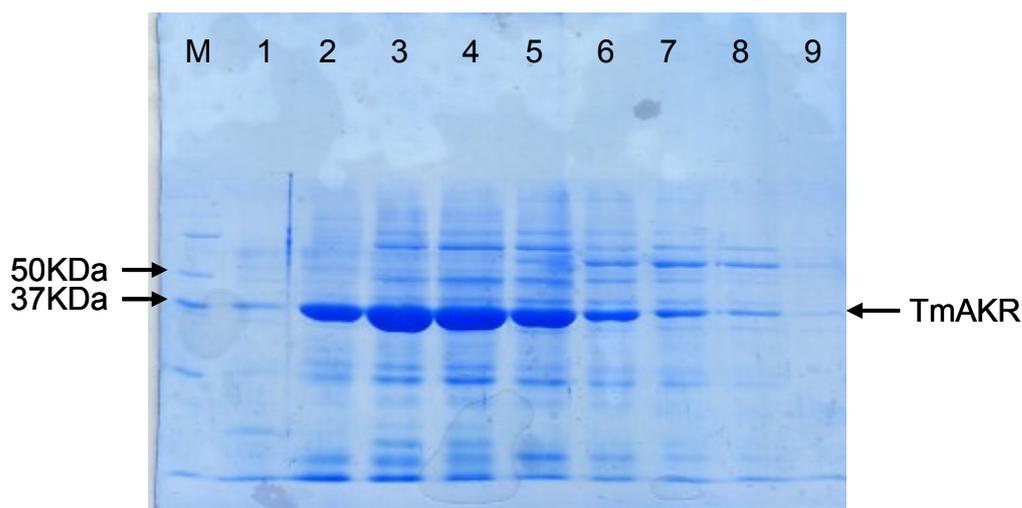
Fig 2.6.   Figure showing the SDS-PAGE analysis of the ion exchange chromatography of TmAKR.

M – Protein molecular weight markers (See 2.2.5.)
Lane 1 – FFQ 220-225ml
Lane 2 – FFQ 225-230ml
Lane 3 – FFQ 230-235ml
Lane 4 – FFQ 235-240ml
Lane 5 – FFQ 240-245ml
Lane 6 – FFQ 245-250ml
Lane 7 – FFQ 250-255ml
Lane 8 – FFQ 255-260ml
Lane 9 – FFQ 270-275ml

The pooled fractions from the ion exchange chromatography were concentrated to 2 mL and loaded onto the Superdex200 gel filtration column. This step does not provide much noticeable purification when analysed by SDS-PAGE (Fig 2.8), but the elution profile (Fig 2.7) shows that higher molecular weight aggregates have been removed providing the high quality and purity of protein required for crystallisation. The peak at 63 mL (240 kDa) contains enzymatic activity and a large proportion of TmAKR and corresponds to twice the molecular weight of the main peak suggesting this could be beginning to form aggregates. Therefore the protein from the peak at 76 mL (112 kDa) was pooled together without that from any other peak. Approximately 12 mg of purified protein was produced from this purification protocol starting with six litres of culture medium.
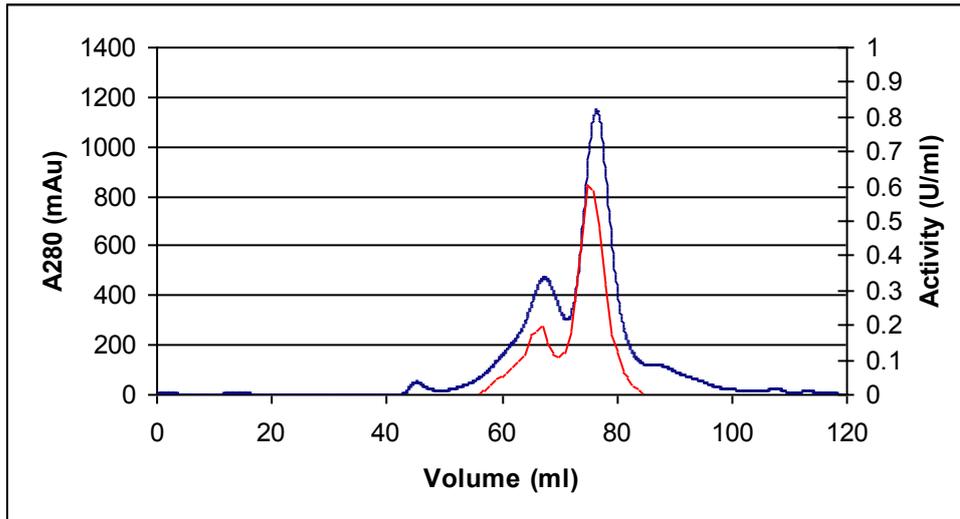
Fig 2.7. Figure showing the A_{280} elution profile (blue) and activity profile (red) of the size exclusion chromatography of the TmAKR protein.
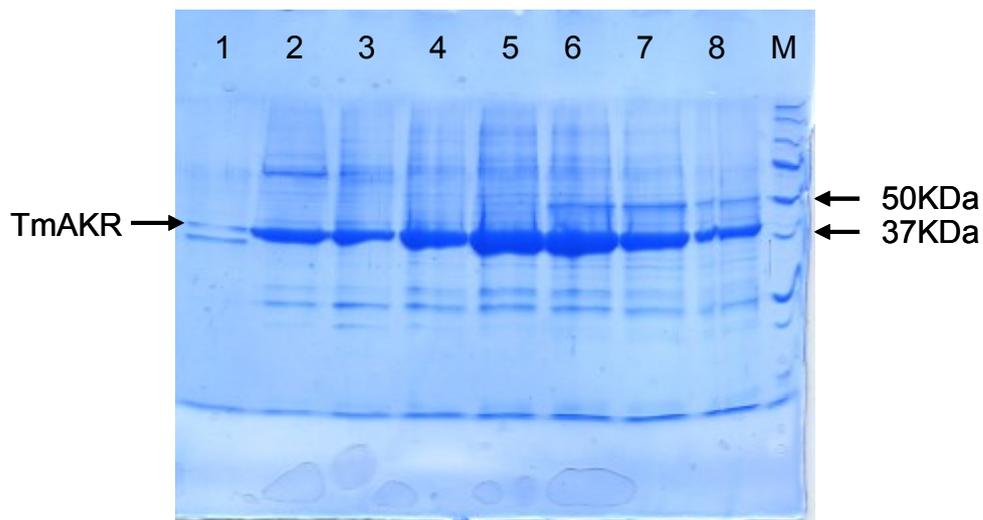
Fig 2.8. Figure showing the SDS-PAGE analysis of size exclusion chromatography of the TmAKR protein.

Lane 1 – GF 40-41ml
Lane 2 – GF 62-63ml
Lane 3 – GF 66-67ml
Lane 4 – GF 68-69ml
Lane 5 – GF 70-71ml
Lane 6 – GF 72-73ml
Lane 7 – GF 74-75ml
Lane 8 – GF 76-77ml
M – Protein molecular weight markers (See 2.2.5)

2.3.3 Bacterioferritin Purification

Early on in developing the purification protocol for TmAKR it was noticed that on some occasions the TmAKR co-purified with an *E. coli* protein in moderate quantities and they were separated by size exclusion chromatography. The purification procedure was identical to that described above for the TmAKR protein except the cells were lysed by sonication prior to heat treatment at 80°C. The difference in the two purification protocols was due to the further development of the TmAKR protocol. The bacterioferritin was purified twice with the protocol described (section 2.2.8) and despite efforts to repeat the purification with identical conditions the bacterioferritin was not found again. At this stage the identity of the bacterioferritin was not known and was not identified as *E. coli* bacterioferritin until solution of the X-ray crystal structure (Chapter 5)

The elution profile of the Superdex 200 gel filtration chromatography of such purifications can be seen in Fig 2.9 and the corresponding SDS-PAGE analysis (Fig 2.10)
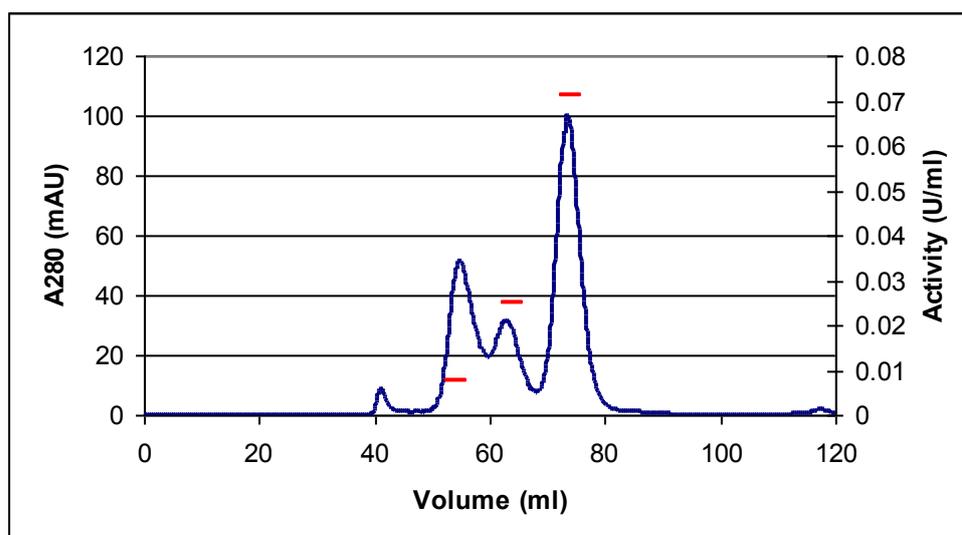
Fig 2.9. Figure showing the A_{280} elution profile (blue) of TmAKR and Bacterioferritin separated by gel filtration chromatography on Superdex 200. The TmAKR activity for pooled fractions is shown as red bars.
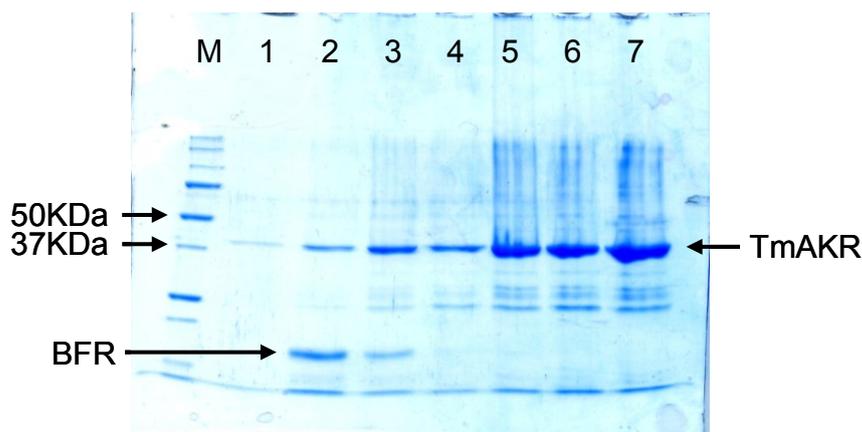
Fig 2.10. SDS-PAGE of size exclusion chromatography from Fig 2.9.

M – Protein markers (see 2.5.5)
Lane 1 – 41-42ml
Lane 2 – 54-55ml
Lane 3 – 58-59ml
Lane 4 – 62-63ml
Lane 5 – 71-72ml
Lane 6 – 73-74ml
Lane 7 – 75-76ml

As can be seen from the SDS-PAGE analysis of the gel filtration all peaks contain the AKR in significant proportions. The peak containing the bacterioferritin eluting at approximately 56 mL corresponds to a molecular weight of 458 kDa and can be seen to contain both the active AKR and another protein with a subunit molecular weight of 15 – 20 kDa. This larger molecular weight peak corresponds to four times the weight of the main AKR peak (76 mL; 112 kDa) with the intermediate peak (63 mL; 220 kDa) corresponding to twice the molecular weight of the main peak and could correspond to progressively higher order aggregates on the route to insoluble aggregation. The AKR may be interacting with a contaminating protein affecting its retention time, or could be due to a higher oligomeric state of the TmAKR protein. Due to the presence of AKR activity, this peak (at 56 mL) was pooled and taken forward to crystallisation trials (Chapter 3). The intermediate peak at 63 mL was not pooled as too little protein was present for crystallisation trials.

3 AKR Characterisation

3.1 Introduction

To evaluate the TmAKR enzyme for use as a biocatalyst in biotransformation reactions it is required to characterise the enzyme with respect to several important properties. It is preferable for an enzyme to be active and stable at elevated temperatures. This allows the process to be performed at higher temperatures, and also gives an indication of the potential long term stability of the enzyme and the possibility of recycling it for reuse. The substrate used in a biotransformation may be only sparingly soluble in aqueous media, therefore an organic co-solvent may be required to solubilise the substrate to increase the enzyme's performance in such a system. The enzyme also needs to have broad substrate specificity so it may be used in a wide range of reactions, and in the case of oxidoreductases the enzyme needs to be high enantioselectivity towards production of chiral secondary alcohols.

The conversion of a carbonyl to the alcohol is referred to here as the forward reaction, and the reverse reaction as the conversion of an alcohol to the corresponding carbonyl.

3.2 Materials and Methods

TmAKR used for characterisation experiments was from purified protein after gel filtration chromatography (Section 2.2.7.3) which was stored at -20°C at 0.5 mg mL^{-1}. Once defrosted, the enzyme concentration was re-estimated by absorption at 280nm. Screening experiments were performed with partially purified TmAKR from the heat lysis (section 2.2.7.1) which was stored at -20°C. All graphs were produced using Sigmaplot 10.0 (SYSTAT software). One unit of activity is defined as one µmole of NADPH reduced in one minute.

3.2.1 Activity Assay

Standard activity assays were performed as previously described in section 2.2.6.

3.2.2 Substrate Screening

Substrate screening was performed in 96 well plates and the change in absorbance at 340nm followed over a period of 20 minutes on a Tecan rainbow 96 monochromatic microplate reader. Substrates were prepared to 30% (v/v for liquids or w/v for solids) in dimethyl sulfoxide (DMSO) in a 96 well microplate containing glass vial inserts (Chromacol). 75 µL of 50mM Imidazole pH 6.5 containing 15mM NADPH was added to each well of a 96 well plate. 5 µL of each substrate was added into the appropriate well, and the reaction started by addition of 20 µL TmAKR (heat treated extract). A list of the substrates tested is shown in table 3.1 and the structures of these are shown in appendix 3.

1,2,3,6-tetrahydrobenzaldehyde	ethyl benzoylacetate
2-chloro-3',4'-dihydroxyacetophenone	propiophenone
dichlorobenzophenone	2-acetyl pyridine
1,4-dibromo-2,3-butanedione	4-acetylbenzonitrile
2-bromoacetophenone	3-hexanone
1,3-dimethylimidazolidinone	cyclopentanone
2-butanone	3-acetylpyridine
cyclohexanone	acetophenone
2-cyclohexen-1-one	2-chloroacetophenone
2-hexanone	3,5-bis(trifluromethyl)acetophenone
4-acetylpyridine	p-anisaldehyde
acetonylacetone	m-anisaldehyde
3-octanone	benzaldehyde
2'-chloroacetopenone	acetone
2'-(trifluromethyl)acetophenone	

Table 3.1. Ketone and aldehyde substrates used for screening assays with TmAKR.

Potential substrates identified from spectrophotometric screening were scaled up and analysed by GC to confirm activity. 1 mL of 1M Tris-HCl pH 7.6 containing 8 mg NADPH, 100 mg D-glucose and 0.28 mg Glucose Dehydrogenase (CDX901, Codexis) were placed in a scintillation vial with 100 µL DMSO, 20 mg substrate and 1 mL TmAKR (heat treated extract). Reactions were incubated at 25°C shaking overnight. 200 µL samples were taken, and reaction products extracted into 1 ml MTBE. Samples were analysed for substrate depletion and product accumulation by GC, using a generic method with a DEX CB column at an initial temperature of 60°C for 2 minutes, ramping to 220°C at 10°C per minute and holding for 2 minutes. Substrate and product identity was confirmed by comparison of retention times for commercially available pure samples.

3.2.3 pH Optimum

Assays were performed in a Shimadzu UV2100 double beam spectrometer at 50°C. A range of assay buffers were prepared, covering the pH range from 5.5 - 11 and adjusted to the appropriate pH at the assay temperature. Each buffer was incubated at

50°C for 5 minutes prior to the assay being carried out. For the forward reaction 960 µL of the assay buffer was placed in a quartz cuvette with 2 µL THBA and 20 µL of 5mM NADPH. For the reverse reaction 960 µL of the assay buffer was placed in a quartz cuvette with 2 µL 3-cyclohexene-1-methanol and 20 µl of 20mM NADP. The assay was started with the addition of 20 µL TmAKR and the change in absorbance at 340nm followed for 5 minutes. Initial rates were estimated from the resulting curve and specific activities calculated using the protein concentration estimated from A_{280} readings and the extinction coefficient. All assays were performed in triplicate.

3.2.4 Temperature Optimum

Assays were performed in a Shimadzu UV2100 double beam spectrometer. 50mM Imidazole pH 6.5 was incubated at the appropriate assay temperature for five minutes prior to the assay being carried out. The buffer was adjusted to the correct pH at the appropriate assay temperature. 960 µl of the assay buffer was placed in a quartz cuvette with 2 µl THBA and 20 µL NADPH (5mM). 20 µL TmAKR was added to start the reaction and the change in absorbance at 340nm followed for 5 minutes. Assays were performed in triplicate and initial rates were calculated from the resulting curve and specific activities calculated using the protein concentration estimated from A_{280} readings and the extinction coefficient. The assay was performed at 30°C and at 5° intervals from 40°C to 70°C. Operational limits of the equipment prevented measurement of activity at higher temperatures.

3.2.5 Temperature Stability

The TmAKR was incubated in a water bath at a range of temperatures (70-90°C) for between 30 and 300 minutes. After the incubation period was over the sample was removed to ice and allowed to cool for 30 minutes prior to being assayed. Assays were performed as per the standard assay described in section 2.2.6. Assays were performed in triplicate and remaining activity calculated with reference to an unincubated sample from the same batch.

3.2.6 Solvent Stability

TmAKR was incubated at room temperature with 5, 20, 40 and 60% of a range of organic solvents (isopropanol, butan-1-ol, acetonitrile and DMSO). The appropriate amount of organic solvent was mixed with 50mM Imidazole pH 6.5 to a volume of 960 µL, 2 µL of THBA and 20 µL of enzyme was added and mixed gently. The

sample was then incubated at room temperature for 30 minutes. After incubation the sample was immediately placed in a quartz cuvette at 50°C for 5 minutes. 20 µL NADPH (5mM) was added to start the reaction and the change in absorbance at 340nm followed for 5 minutes. Assays were performed in triplicate and remaining activity calculated with reference to an unincubated sample from the same batch.

3.2.7 Michaelis-Menten Kinetics

The kinetics of the TmAKR catalysis of THBA was investigated. Assays were carried out as described in section 2.2.6, except the substrate concentration was varied from 0.44µM to 22µM. 20µl of 0.204 mg mL^{-1} protein was used per assay, and assays were performed in triplicate. Initial rates were calculated and Sigmaplot10.0 (SYSTAT software) used to calculate K_m and V_{max} values by linear regression.

3.3 Results and Discussion

3.3.1 Substrate Screening

Substrates used in screening reactions were chosen to provide a range of varying chain length, branched, non-branched and aromatic substrates. Acetone was specifically chosen to investigate substrate coupled cofactor recycling (section 1.2.1). Spectrophotometric screening identified nine substrates which should be investigated further as potential substrates. These were benzaldehyde, 3,5-bis(trifluoromethyl)acetophenone, 2'-(trifluoromethyl)acetophenone, 2-cyclohexen-1-one, THBA, 4-acetylpyridine, dichlorobenzophenone, acetonylacetone and para-anisaldehyde. These substrates identified from initial spectrophotometric screening were subject to further screening reactions and analysis by GC. Three substrates were identified to exhibit activity with the TmAKR enzyme (Fig 3.1). The three substrates identified (Fig 3.2) were benzaldehyde, THBA and para-anisaldehyde. All substrates are aldehydes containing a bulky phenyl group or a partially hydrogenated phenyl group. No activity with acetone was found, and therefore substrate coupled cofactor regeneration would not be an option unless the activity could be engineered to accommodate acetone/isopropanol. The enzyme has shown specificity for aldehydes and not for ketones, limiting the enzyme to the production of primary alcohols. Therefore any product would not be able to possess a chiral centre at the carbonyl carbon, but chiral centres may be present at other locations within the substrate or product. One of the substrates THBA, possesses a chiral centre at C1 (Fig 3.2) due to the partially hydrogenated nature of the aromatic ring. GC analysis was unable to resolve the different enantiomers of the substrate and product; therefore it was not possible to determine if the enzyme was selective towards one enantiomer of this substrate. The specificity of the enzyme towards aldehydes and the lack of activity with ketones suggest there may be steric hindrance in the corresponding part of the enzyme precluding any activity towards ketones and secondary alcohols.
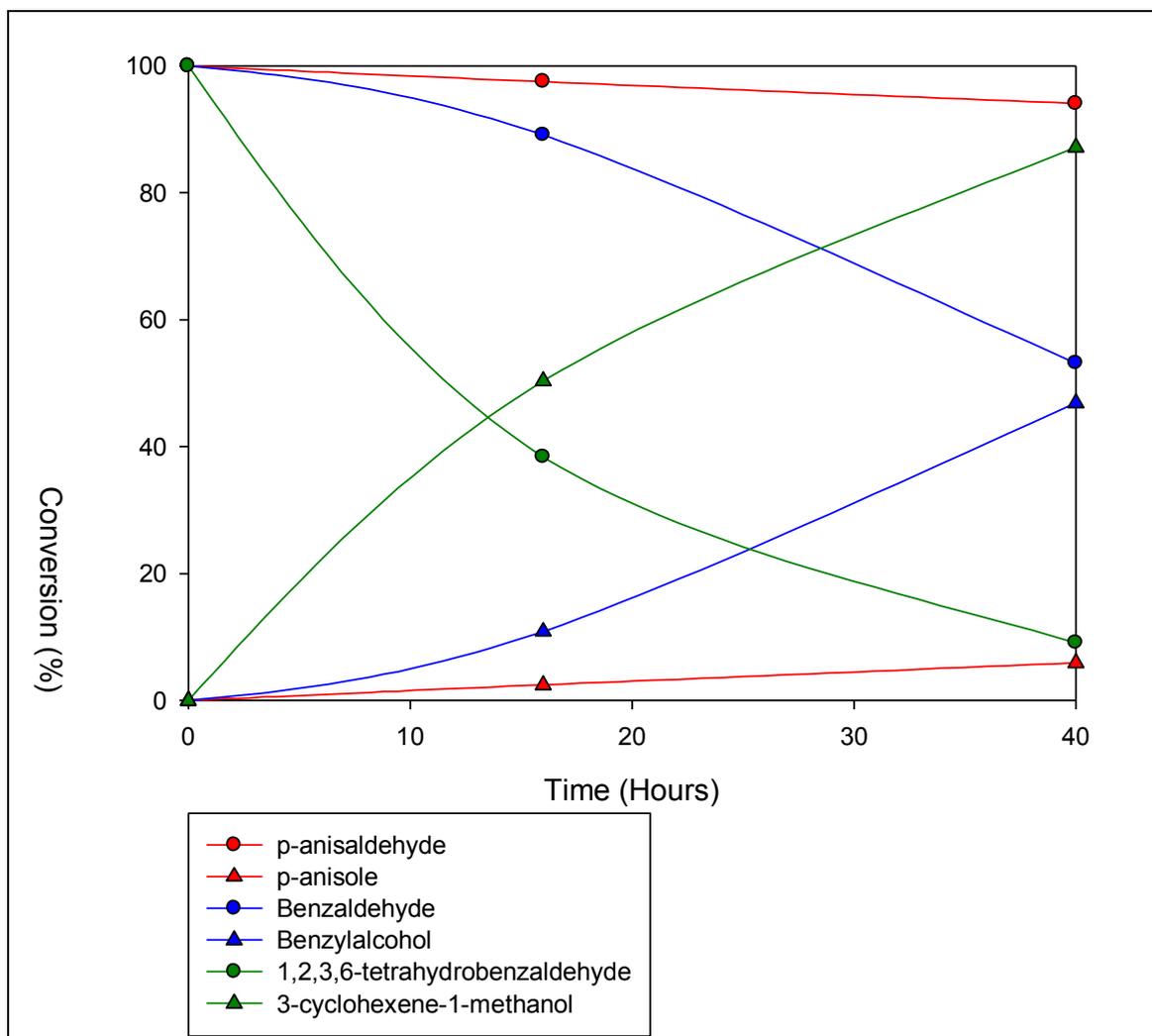
Fig 3.1. Confirmation of activity of TmAKR towards potential substrates from GC analysis, showing conversion of the three substrates identified from screening to the corresponding alcohol product. The percentage abundance is measured relative to the starting abundance of the substrate carbonyl.

Benzaldehyde   1,2,3,6-tetrahydrobenzaldehyde

para-anisaldehyde
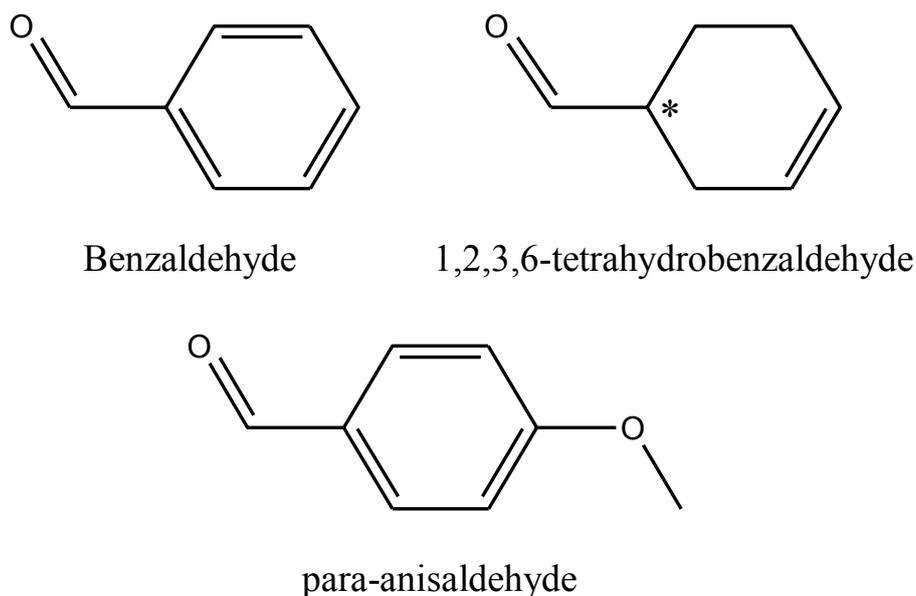
Fig 3.2. Substrates to which TmAKR has been shown to be active. * - Chiral centre

3.3.2 pH Optimum

The pH optima for the forward and reverse reactions were investigated and are shown in Fig 3.3 and Fig 3.4 respectively. The forward reaction shows preference towards activity at lower pH values and exhibits a pH optimum around pH 6.5 in 50mM imidazole buffer. Activity in phosphate buffer was found to be reduced, and may be due to interference with cofactor binding, as the buffer may be able to occupy sites associated with the $NADP^+$ phosphate binding. The activity of TmAKR in imidazole buffer appears to be shifted towards higher pH values in relation to the activity in phosphate buffer. This shift in activity may be due to some unknown specific interaction of imidazole with the enzyme. The reverse, carbonyl to alcohol, reaction appears to have a pH optimum around pH 8.5 in BICINE buffer. The activity observed is marginally above the lower limit of detection for this method, therefore the pH optimum could not be accurately determined and the true pH optimum may be within the pH 8.5 to pH 10 region. The reverse reaction shows much lower activity (20 fold less) than the forward reaction, suggesting this enzyme is only utilised in alcohol to carbonyl metabolism under physiological conditions. Due to this low activity for the reverse reaction and the errors associated in measuring this, all further characterisation concentrated on the forward, carbonyl to alcohol reaction.
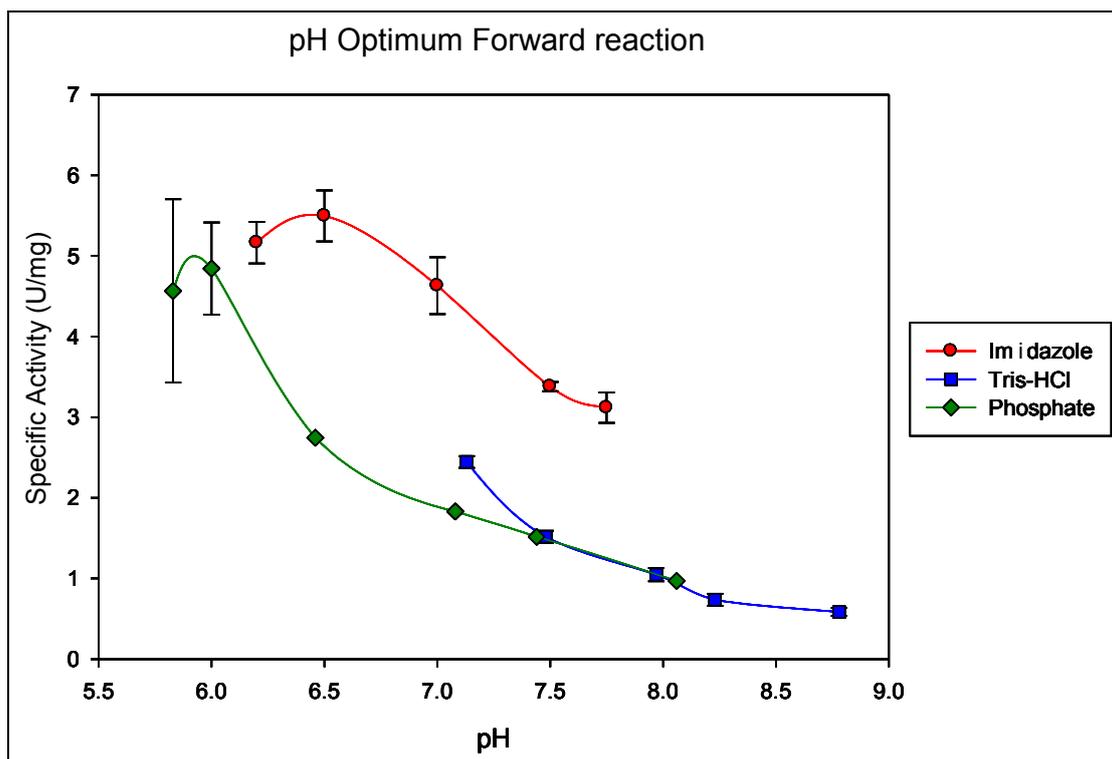
Fig 3.3. The pH optimum of the forward reaction catalysing the conversion of alcohol to carbonyl.
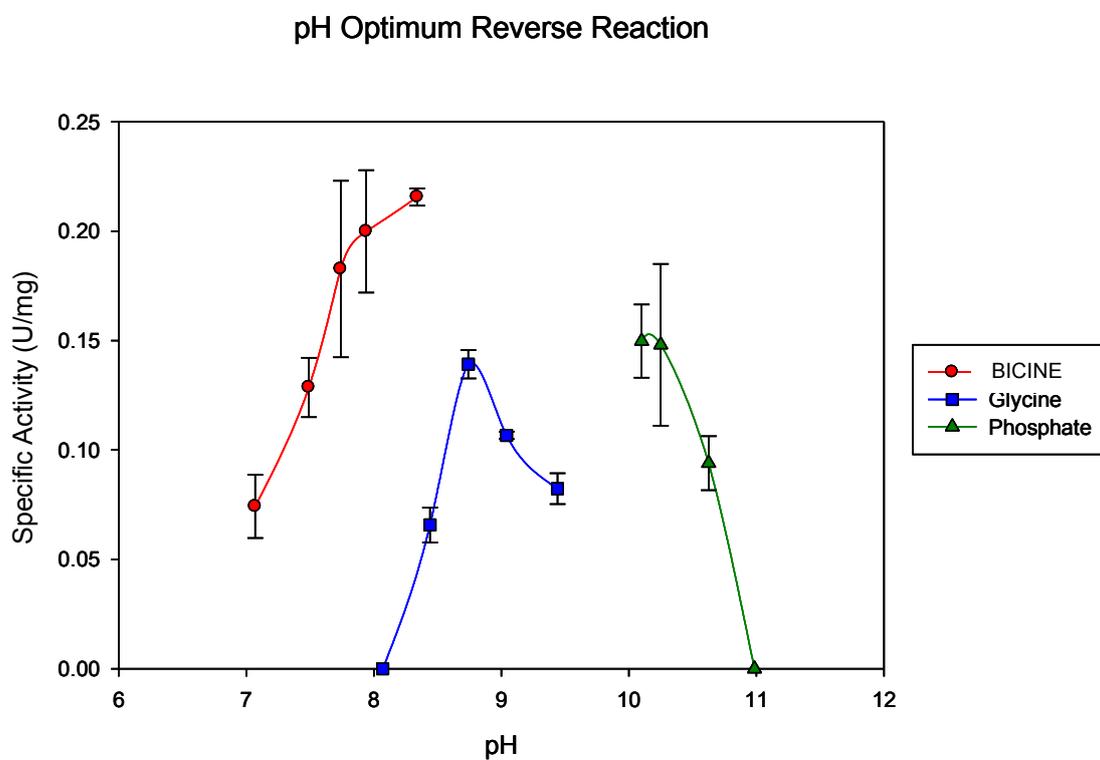
Fig 3.4. The pH optimum of the reverse reaction catalysing the conversion of carbonyl to alcohol. Note differing scale from Fig 3.4.

3.3.3 Temperature Optimum

The effect of temperature on the rate of reaction was investigated from 30°C upto the experimental limits of the equipment. The results (Fig 3.5) show a consistent increase in activity over the range studied and the reaction does not reach a maximum over this range. There is a noticeable increase in the error observed at higher temperatures due to the increased rate of degradation of NADPH cofactor at higher temperatures. Therefore all characterisation assays were performed at 50°C.
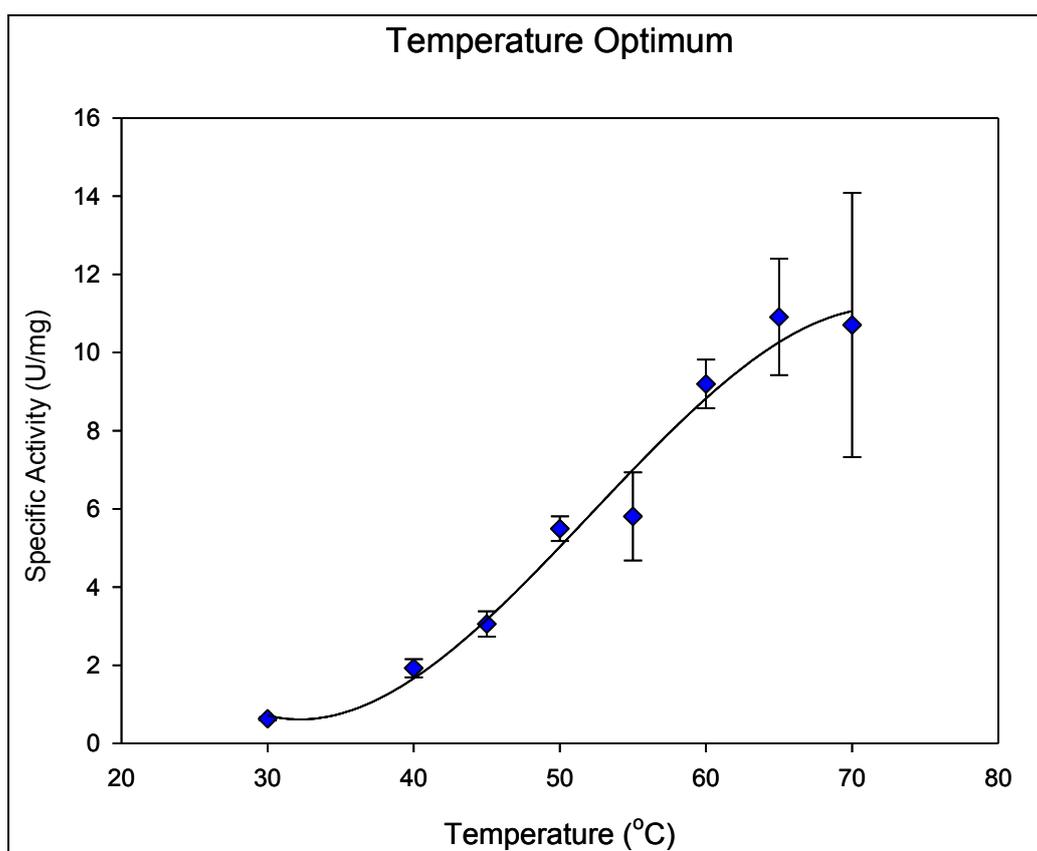
Fig 3.5. The temperature optimum for the alcohol to carbonyl reaction of TmAKR.

3.3.4 Temperature Stability

The thermal inactivation of the TmAKR enzyme was investigated over a period of 5 hours and with temperatures ranging from 70°C to 90°C. At 70°C there is very little activity lost over the first 1 or 2 hours, with only 20% being lost after 5 hours. Between 70°C and 80°C there is only a small increase in the amount of enzyme activity lost, but above this temperature there is a large decrease in the stability of the enzyme. At 85°C almost all the activity has been lost after 5 hours compared with just 40% loss at 80°C for the same time. At 90°C all the activity has been lost after

only 90 minutes. These results suggest that a vital stabilising element is lost at temperatures above 80°C causing the rate of thermal inactivation to increase. The TmAKR is stable at temperatures up to 80°C for long periods and therefore would be potentially useful in biocatalytic processes as it would be able to withstand the increased temperatures of process conditions.
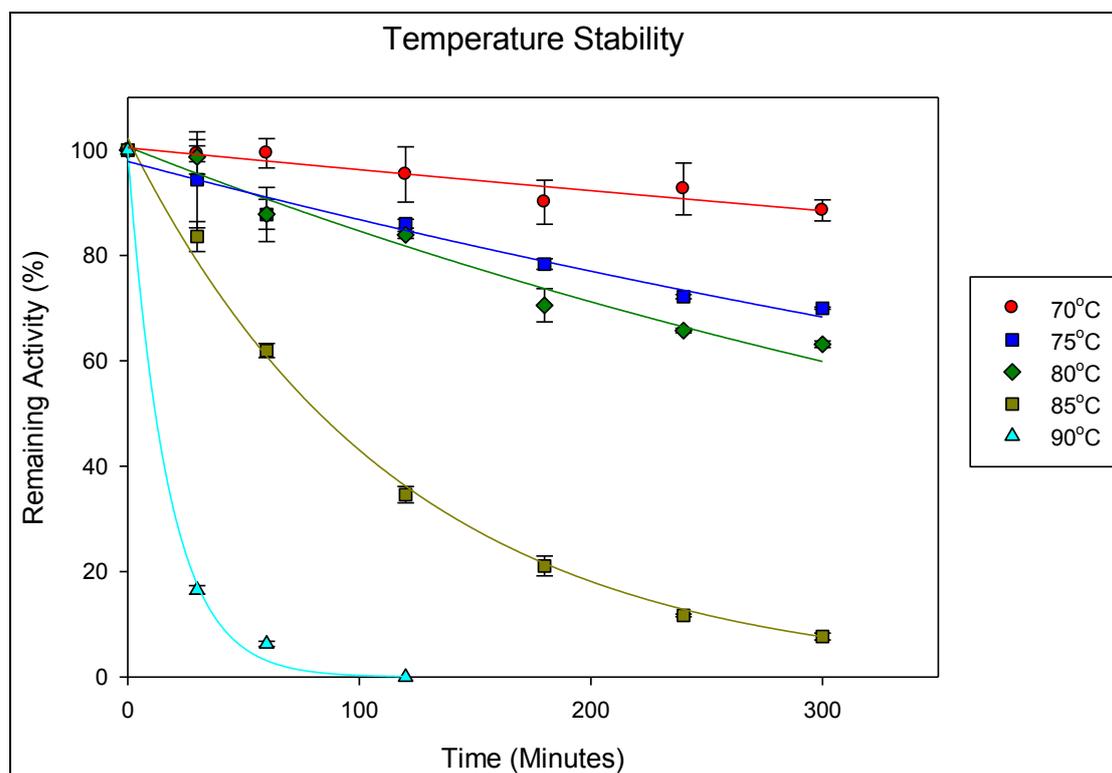
Fig 3.6. The temperature stability of TmAKR over a range of incubation times and temperatures.

3.3.5 Solvent Stability

The ability of the enzyme to catalyse reactions in the presence of organic solvents is important if the enzyme is to be used as a biocatalyst since many substrates will require a co-solvent due to low solubility in aqueous media. The results of these experiments are shown the figure 3.7.
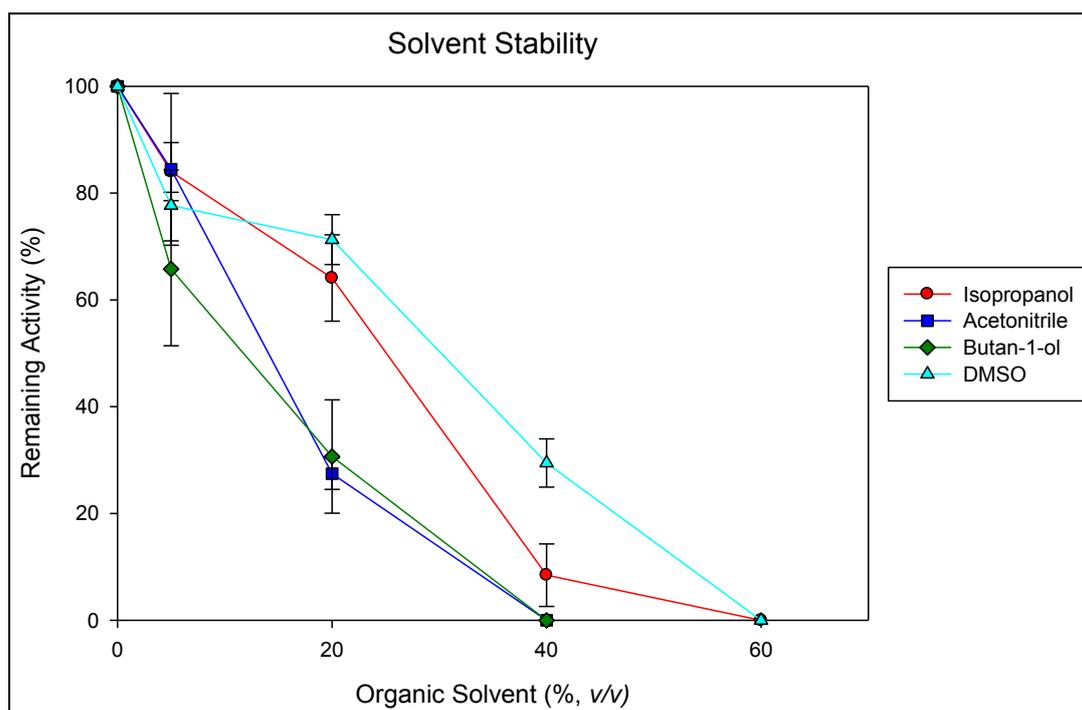
Fig 3.7. Stability of TmAKR with respect to a variety of organic solvents over a range of concentrations.

The enzyme showed no activity above 60% in any organic solvent tested probably due to loss of water from the structure of the enzyme leading to denaturation of the protein. By comparing the stability of the enzyme in each solvent with the polarity index (Snyder, 1974) a general trend can be seen for higher stability in solvents with higher polarity index values (Table 3.2). The only deviation from this trend being isopropanol, which performs favourably but has the lowest polarity index.

Solvent	Polarity Index	Activity Rank
Butan-1-ol	4.0	5
Acetonitrile	5.8	4
Isopropanol	3.9	3
DMSO	7.2	2
Water (no solvent)	10.2	1

Table 3.2. Table showing the polarity index of solvents tested, ranked in order TmAKR activity.

3.3.6 Michaelis-Menten Kinetics

A simple one substrate one product enzyme catalysed reaction may be described as in Eq. 3.1, where free enzyme (E) and substrate (S) form the complex ES (k_1), which may either dissociate to give free enzyme and substrate (k_{-1}), or may undergo catalysis to give free enzyme and product (k_2).

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} E + P \qquad \text{Eq. 3.1}$$
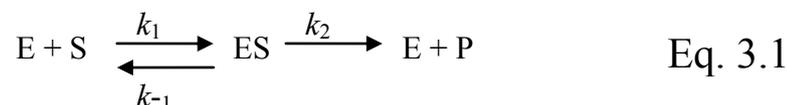
For a reaction which obeys Michaelis-Menten kinetics, the initial rate of reaction (V_0) may be described by the Michaelis-Menten equation (Eq. 3.2). Where V_{max} is the maximum rate of reaction, $[S]_0$ is the initial substrate concentration, and K_m is the Michaelis constant. A plot of V_0 against [S] will produce a rectangular hyperbola for an enzyme reaction obeying Michaelis-Menten kinetics.

$$V_0 = \frac{V_{max}[S]_0}{K_m + [S]_0} \qquad \text{Eq. 3.2}$$

The Michaelis constant (Eq. 3.3) describes the breakdown to the ES complex and is the substrate concentration where the reaction velocity is equal to $V_{max}/2$. At high substrate concentrations the product formation step (k_2) is the rate limiting step, here k_2 is small and the Michaelis constant (K_m) approximates to k_{-1}/k_1, the dissociation constant (k_d) for the ES complex. At lower substrate concentrations formation of the ES complex becomes rate limiting and the rate becomes directly proportional to the substrate concentration. A low K_m value signifies that the enzyme has a high affinity for the substrate ($k_1 > k_{-1}$).

$$K_m = \frac{k_{-1} + k_2}{k_1} \qquad \text{Eq. 3.3}$$

The catalytic constant of an enzyme (K_{cat}), gives the number of molecules of substrate turned over per unit time, and is defined as;

$$K_{cat} = \frac{V_{max}}{[E]_T}$$

Eq. 3.4

Where E_T is the total enzyme concentration. If the reaction obeys Michaelis-Menten kinetics then K_{cat} is equal to k_2. The K_{cat}/K_m value may be used as a measure of an enzymes catalytic efficiency, allowing us to directly compare the performance of different enzymes.

The K_m and V_{max} values for the TmAKR enzyme have been calculated for the turnover of 1,2,3,6-tetrahydrobenzadlehyde (Fig. 3.8). K_{cat} and K_m/K_{cat} were also calculated (Table 3.3). In comparison to the most closely related (by sequence identity) AKRs, AKR11A and AKR11B from *Bacillus subtilis* (Ehrensberger and Wilson, 2004) and AKR11C from *Bacillus halodurans* (Marquardt *et al*., 2005), the TmAKR enzyme has a much lower K_m (μM range rather than mM) but has a slower turnover of approximately 0.3 S^{-1} rather than 1-5 S^{-1} for the other AKRs with the substrate benzaldehyde. The K_{cat}/K_m value for TmAKR is in the region of 100 fold higher than that for AKR11B, meaning that the TmAKR enzyme is more efficient than its related enzyme. Therefore this enzyme appears to be catalytically slower, but has a higher affinity for its substrate making it overall more efficient than the related enzymes.
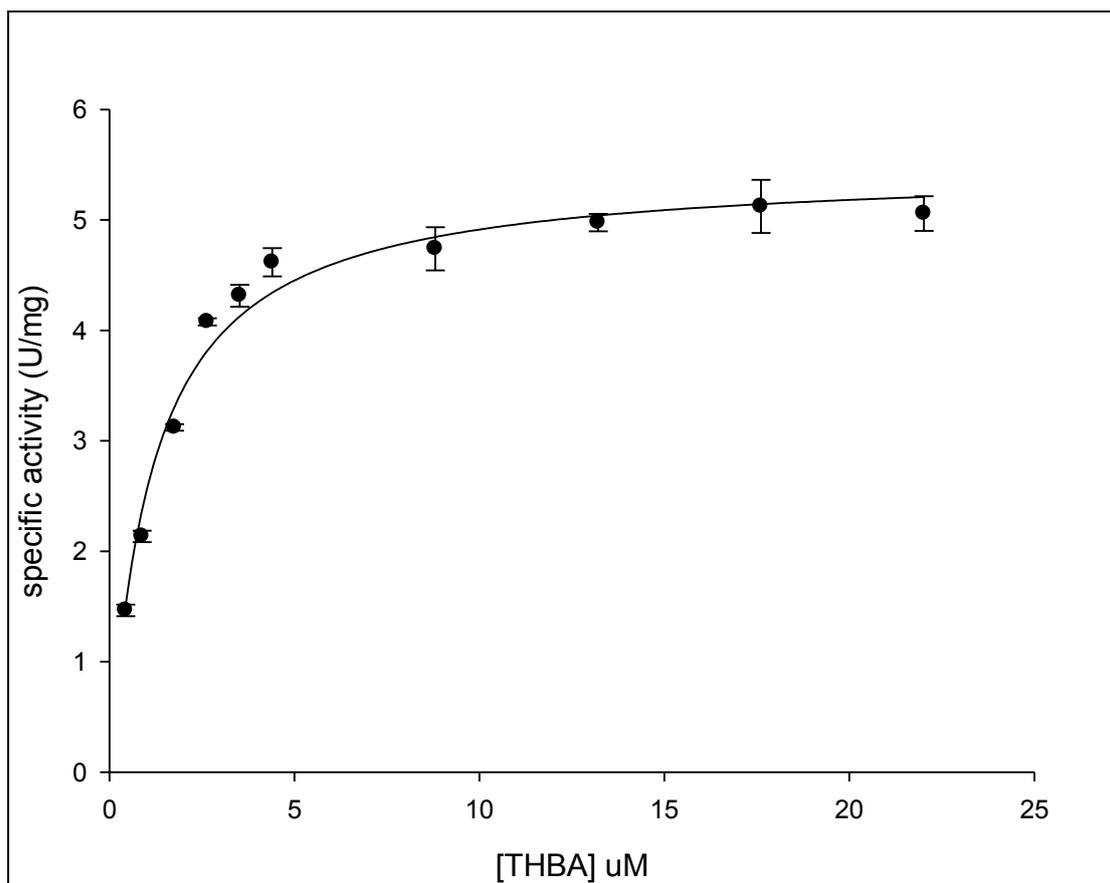
Fig. 3.8. A plot of initial velocity against substrate concentration for TmAKR, showing rectangular hyperbola from linear regression used to calculate K_m and V_{max}. Error bars show 1 standard deviation. THBA – 1,2,3,6-tetrahydrobenzaldehyde

	V_{max} (μmoles min^{-1} mg^{-1})	K_m (μM)	K_{cat} (S^{-1})	K_{cat}/K_m (S^{-1} μM^{-1})
THBA	5.48	1.1561	0.3	0.266

Table 3.3. Summary of kinetic parameters of the TmAKR enzyme with respect to THBA (1,2,3,6-tetrahydrobenzaldehyde).

4 Crystallisation

4.1 Introduction

To produce protein structures by X-ray diffraction methods requires the production of high quality diffracting crystals of the macromolecule of interest. These crystals are a highly ordered three dimensional arrangement of the protein molecules in the solid phase. The transition of the protein from solution to this highly ordered solid phase is dependant upon a number of factors and needs to be controlled carefully to ensure crystals form with sufficiently high order and to the required size. Crystal growth is dependant upon the concentration of all of the ingredients in the crystallisation solution (protein, buffer, precipitants and additives), purity of the protein, solution pH, precipitant type, incubation temperature, and the method of crystallisation used (microbatch, vapour phase diffusion).

4.1.1 Crystallisation Techniques

There are a number of ways to produce diffraction quality crystals but two main methods have been used here and will be focused on. In vapour diffusion (Fig 4.1) a small volume of the protein is mixed with a precipitant solution and placed either on a micro-bridge (sitting drop) or on an inverted cover slip (hanging drop). In this method each well contains a reservoir of the precipitant solution and is sealed with a cover slip and grease. The concentration of protein and precipitant in the drop slowly increases over time via diffusion of water in the vapour phase until both the drop and reservoir reach equilibrium.
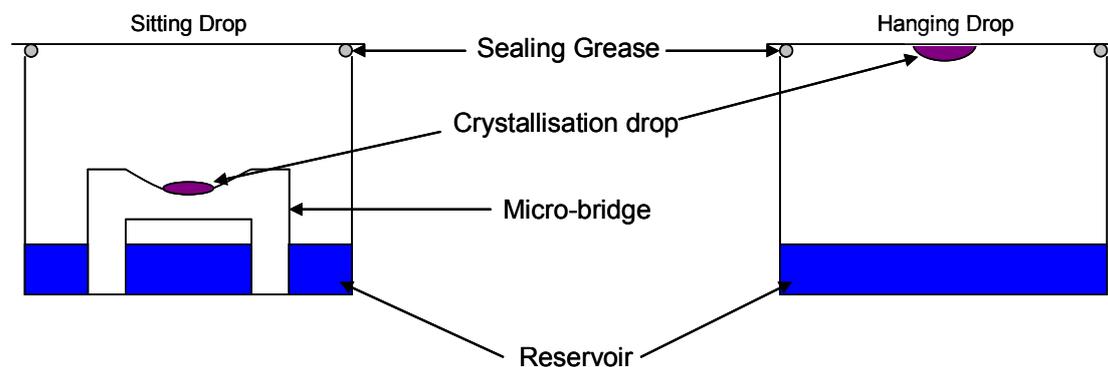
Fig 4.1. The vapour phase diffusion method of protein crystallisation, showing sitting drop and hanging drop techniques. The protein solution either sits on a micro bridge or is suspended by surface tension on a cover slip.

Alternatively with the microbatch method (Fig 4.2) a drop of protein is placed in a well with the precipitant and covered in paraffin oil. This simple to setup experiment provides the protein and precipitant drop at its final concentration from the start, but does not give any significant concentration of the drop and is therefore not useful for sparse matrix screening. By adding silicone oil to the paraffin oil (a mixture often referred to as Al's oils) the evaporation rate can be increased so a similar concentration effect to vapour phase diffusion is achieved (D'arcy, 1996), allowing screening to be performed more successfully.
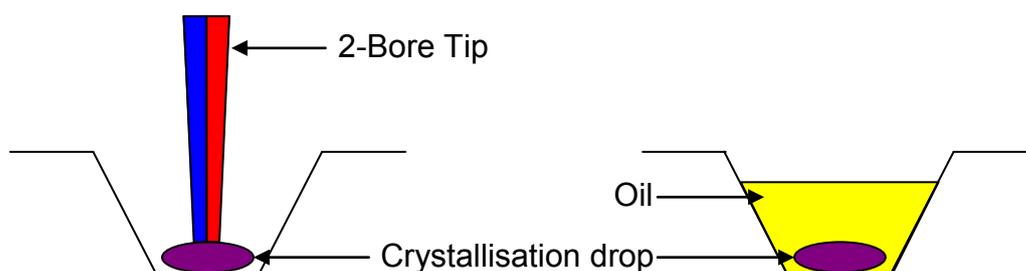
Fig 4.2. Microbatch crystallisation. In each condition both protein and precipitant are pipetted simultaneously by a 2 bore tip. The dispensed drop is then covered with paraffin oil.

The phase diagram (Fig 4.3a) is important in understanding what happens during crystallisation and how to alter the crystallisation conditions to achieve better crystals. When a solution is under-saturated all the components are fully soluble and no crystals will grow. If crystals are added to this phase they will usually shrink or completely disappear. To produce crystals it is necessary to have a supersaturated solution. If however the solution is too highly supersaturated then precipitation will occur and crystals are unlikely to grow. To begin crystallisation the drop must be within the nucleation zone. This is at a high enough concentration to overcome the energy barrier to formation of a small cluster of protein molecules which is the nucleus from which the crystal will grow. In the nucleation zone crystals will also grow as well as nucleate, but once a small number of growing crystals have formed then it is favourable to leave the nucleation zone and enter the metastable zone where crystals will only grow and no new nuclei will form. This allows the maximum amount of protein to be available for crystal growth rather than for further nucleation. This transition from nucleation to growth will usually occur naturally due to protein

moving from the soluble phase into crystals (solid phase) reducing the protein concentration and the nucleation rate (Fig 4.3b,c).
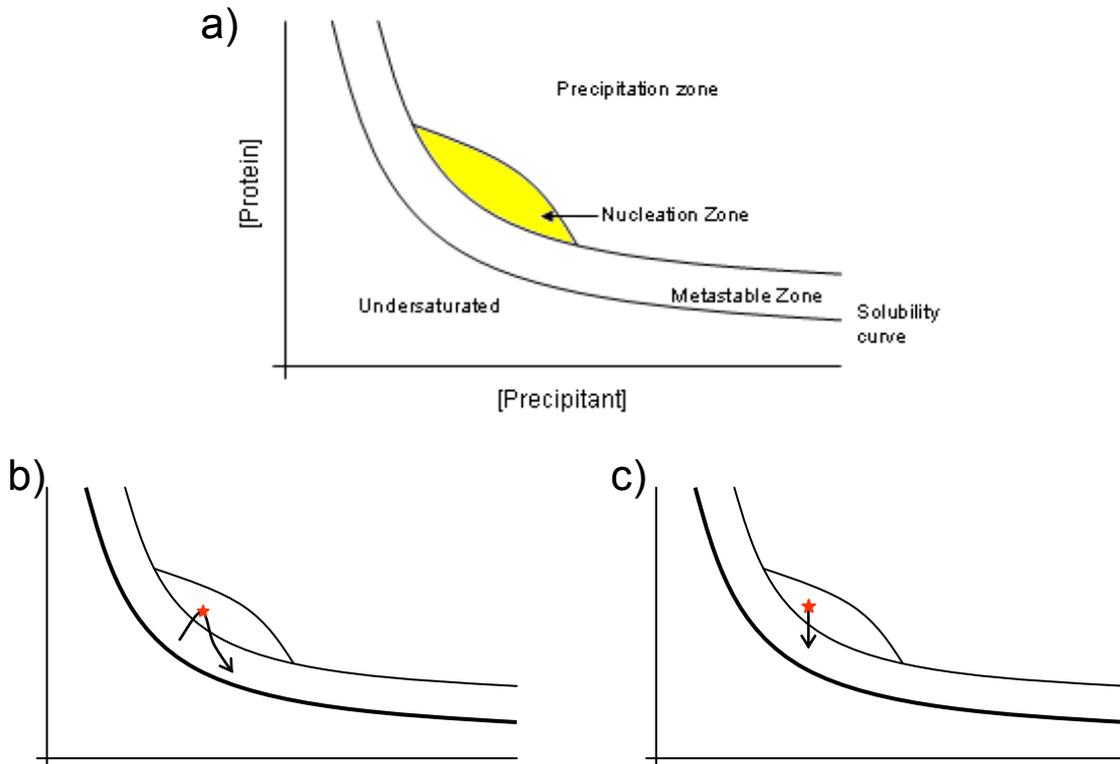
Fig 4.3. Figure showing phase diagrams for protein crystallisation. a) Phase diagram showing solubility curve and undersaturated, metastable, nucleation and precipitation zones.   b) Effect of vapour phase diffusion, initially crystallisation drop is in metastable zone, concentration of the drop increases until nucleation occurs (red star) where protein concentration begins to drop and passes back into metastable zone and growth continues.   c) In microbatch no significant concentration occurs, starting condition must give nucleation for growth to occur, protein concentration drops after nucleation into metastable zone. By using microbatch with 'Al's oils' a concentration effect similar to b) occurs.

4.2 Materials and Methods

4.2.1 Sample Preparation

Pure protein (>95%) from gel filtration chromatography was pooled together and concentrated to the required concentration (10mg ml^{-1} for vapour phase diffusion, 20 mg mL^{-1} for microbatch) using a Vivaspin centrifugal concentrator (Sartorius, UK) with a 30 kDa molecular weight cut off. The sample was centrifuged at 13,000g for 10 minutes to remove any precipitates, prior to crystallisation.

4.2.2 Crystallisation methods

4.2.2.1 Sitting Drop Vapour Phase Diffusion

Vapour phase diffusion crystallisation was carried out in 24-well plates (Greiner) with a microbridge (Molecular Dimensions Ltd) positioned in each well. 500 μL of reservoir solution was placed in each well, 1 μL of protein solution was placed on the microbridge and mixed with 1 μL of reservoir solution by pipetting. Each well was sealed with grease and a cover slip (Molecular Dimensions Ltd). Plates were incubated at 20°C and investigated regularly to check for formation of protein crystals

4.2.2.2 Microbatch

Crystallisation by the microbatch method was carried out using an ORYX-6 robot (Douglas Instruments). For crystallisation trials each precipitant solution is transferred by a 2-bore tip from a stock plate and is pipetted simultaneously with an equal volume of protein solution. After pipetting each condition, 'Al's oil' (1:1 silicone oil: paraffin oil) was placed on top of the drop.

Optimisation experiments were designed using X-step software (Douglas instruments). Optimisation experiments were performed using a 5-bore tip with each channel containing a different ingredient (protein, 2 buffer, precipitant, water). The ingredients were pipetted into each well and mixed with one stir of the tip and 'Al's oil' placed on top.

4.2.3 Crystallisation Trials

4.2.3.1 Bacterioferritin

Protein at 10 mg mL^{-1} (section 4.2.1) was used for crystallisation trials using the sitting drop vapour phase diffusion method (as described 4.2.2.1). Reservoir solutions

were set up over a range of ammonium sulfate concentrations (50-80% saturation) each containing 20mM Tris-HCl pH 7.5 and 0.1M NaCl, 1 µL of protein solution was mixed with 1 µL reservoir solution on a microbridge and the well sealed. Plates were then incubated at 20°C.

4.2.3.2 TmAKR

Crystallisation trials using the sitting drop vapour diffusion method were performed using protein at 10 mg mL^{-1} (section 4.2.1). Trials were performed using the MDL1 and MDL2 structure screens (Molecular Dimensions Ltd) (Appendix 4). 1 µL of protein was mixed with 1 µL of reservoir solution on a microbridge and the wells sealed and incubated at 20°C.

Microbatch crystallisation trials were performed with 20 mg mL^{-1} protein (section 4.2.1) using pHclearI, pHclearII and PEG1 structure screens (Qiagen) (Appendix 4). 1µl of protein was mixed with 1 µL of precipitant solution in a 96 well Douglas microbatch plate (Douglas Instruments), covered with a 1:1 mixture of silicone oil and paraffin oil and incubated at 20°C.

4.2.4 Crystallisation Optimisation

Where crystallisation trials yielded crystals, conditions around these were investigated to improve crystal quality. Optimisation experiments were designed using the X-step software (Douglas Instruments). Stock buffer solutions were made to 1M concentrations, salt precipitants were made to saturation and PEG precipitants were made to 50% (w/v). All optimisations were performed by the microbatch method.

4.3 Results and discussion

4.3.1 Crystallisation Trials

4.3.1.1 Bacterioferritin

Bacterioferritin crystals grew overnight across the whole range of ammonium sulfate concentrations using the vapour phase diffusion method. No further optimisation was required. The best crystals grew in 60% ammonium sulfate (Fig 4.4) to a size of approximately 0.125x0.125x0.125mm and exhibited a bright red colour.
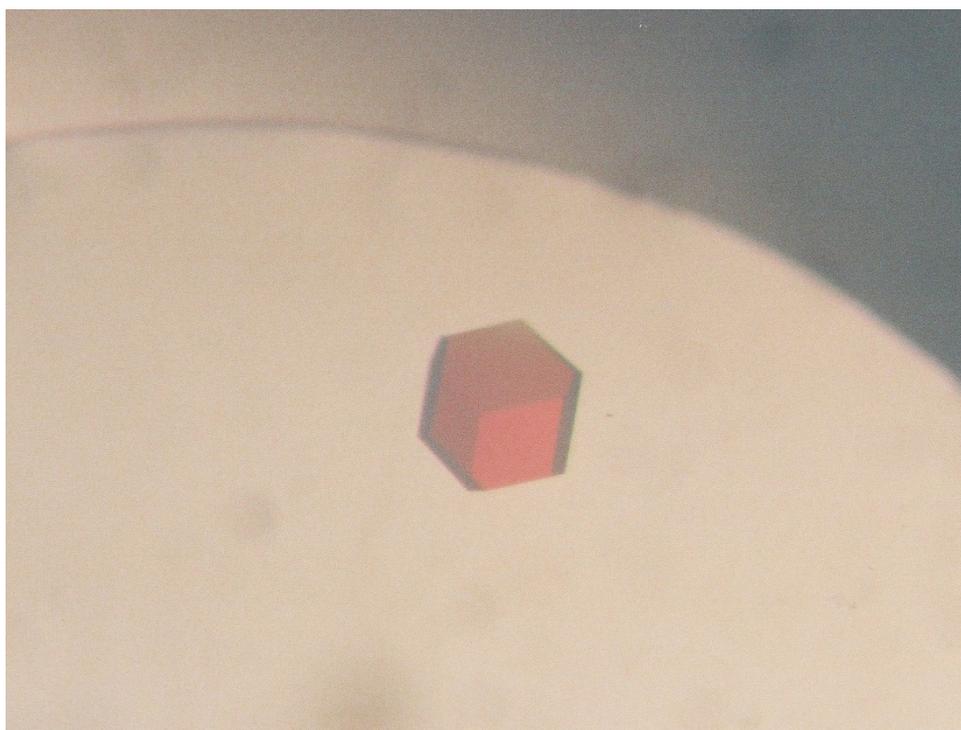
Fig 4.4. A single crystal of *E. coli* bacterioferritin grown in 60% ammonium sulfate, 20mM Tris-HCl pH 7.5, 0.1M NaCl.

4.3.1.2 TmAKR

TmAKR crystals grew over a 2 month period from the pHclearI (Qiagen) structure screen. Crystals (Fig 4.5) were observed in 3 conditions (numbers 32, 55 and 56), 10% (w/v) PEG6000 0.1M Citric acid pH 5.0 (Fig 4.5a) and 1.6M ammonium sulfate 0.1M citric acid pH 5.0 (Fig 4.5b) and pH 4.0 (Fig 4.5c).

All crystals were found in citric acid at either pH 4 or 5, those grown from ammonium sulfate were found as single crystals and the ones from PEG6000 grew as needles. The crystals grown at pH 4 in ammonium sulfate grew under a heavy brown precipitate suggesting this low pH may not be a favourable environment for the

protein, however crystals were of a similar morphology to those grown in similar conditions at pH 5, suggesting the pH range between these two conditions may be worth investigating.
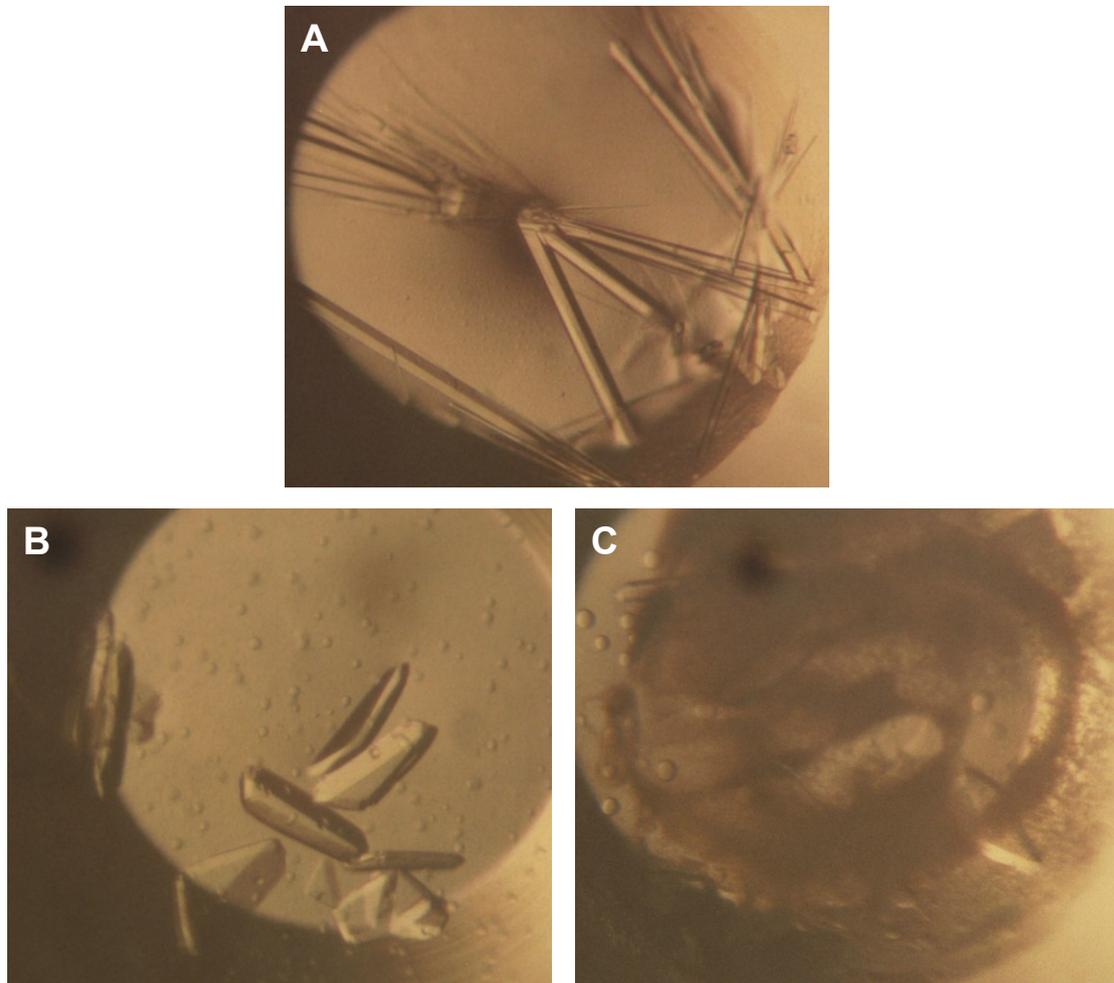
Fig 4.5. Figure showing crystals of TmAKR obtained from microbatch screening. Screening solutions were a) 0.1M citric acid pH 5.0, 10% (w/v) PEG6000 b) 0.1M citric acid pH 5.0, 1.6M ammonium sulfate c) 0.1M citric acid pH 4.0, 1.6M ammonium sulfate.

4.3.2 Crystallisation Optimisation

4.3.2.1 TmAKR

Optimisation of crystal growth conditions were performed in microbatch and experiments designed around those conditions identified from screening using the Xstep optimisation software (Douglas Instruments). The pH, precipitant concentration, and protein concentration were all investigated to improve crystal quality. The ammonium sulfate grown crystals were optimised to produce the best crystals from 0.9M ammonium sulfate, 50mM citric acid pH 5.1 with 7mg ml^{-1} protein (Fig 4.6), all optimised crystals grown in ammonium sulfate grew in a different habit, preferring a more circular or half moon appearance to the triangular form discovered from screening.
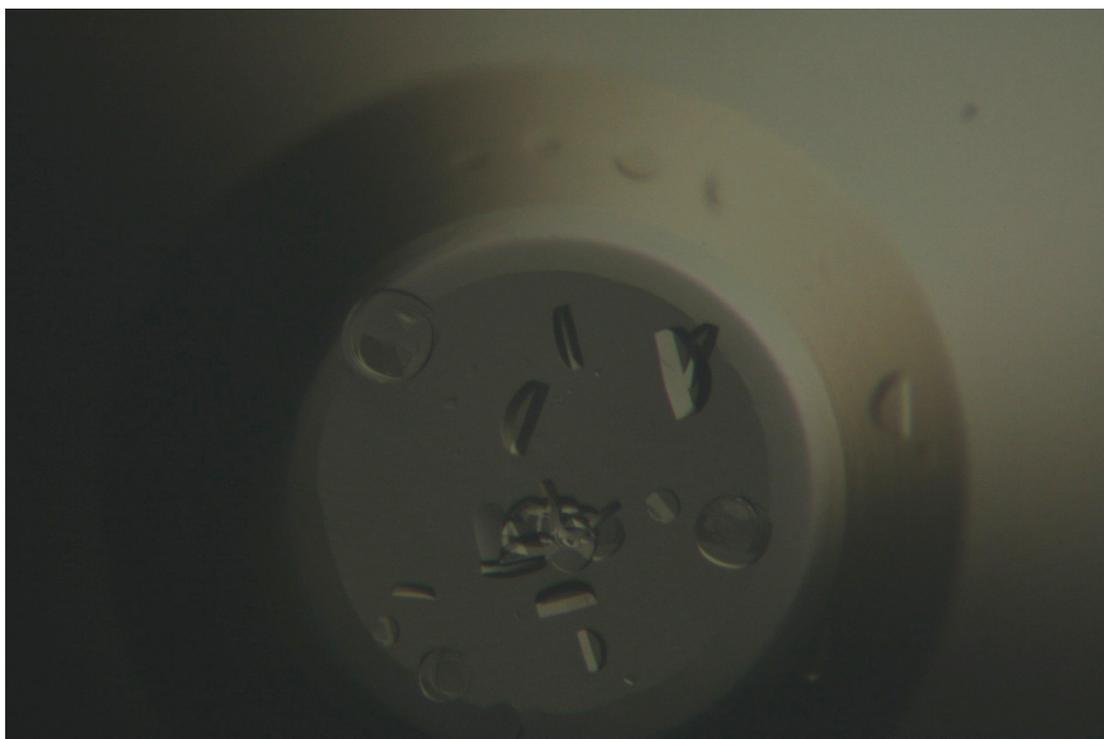
Fig 4.6. Optimised crystals of TmAKR grown in 50mM citric acid pH 5.1, 0.9M ammonium sulfate, 7 mg mL^{-1} protein.

Using these optimised crystals as a starting point it was attempted to co-crystallise in the presence of 1.2 molar equivalents of NADPH to provide a holo structure of the protein. Optimisation was carried out starting at the final conditions from previous optimisation by once again varying the pH, precipitant and protein concentrations.

The buffer composition was also changed to acetic acid to investigate this additional variable. Another crystal form was discovered after changing the buffer to acetic acid (Fig 4.7). The best crystals of this type were grown from 1.6M ammonium sulfate, 50mM acetic acid pH 4.5 with 7.5 mg mL^{-1} protein. These crystals grew as large thin plates with maximum dimensions 0.5 x 1 x 0.01 mm and appeared to look more like the original ammonium sulfate crystals from screening (Fig 4.5B) rather than the optimised crystals (Fig 4.6).

Fig 4.7. Further optimised crystal of TmAKR grown in 1.6M ammonium sulfate, 50mM acetic acid pH 4.5, 7.5 mg mL^{-1} protein and 1.2 Molar equivalents NADPH.

The needle like crystals discovered in screening (Fig 4.5A), were also optimised by varying the pH, protein and precipitant concentrations and altering the buffer composition. These crystals proved harder to grow consistently and to a practical size for X-ray diffraction. Long needles (Fig 4.8 A) were grown but did not produce good faces (indicated by red circle), and become over dried by the time of harvesting. Slightly shorter but thinner needles were also grown (Fig 4.8 B). These crystals could be used but needed to be separated from the clusters of needles growing off them. In certain conditions the nucleation rate had increased (Fig 4.8 C, D) leading to many smaller crystals and more clusters of crystals. These drops however do contain a few single short needles which could be used for X-ray diffraction.
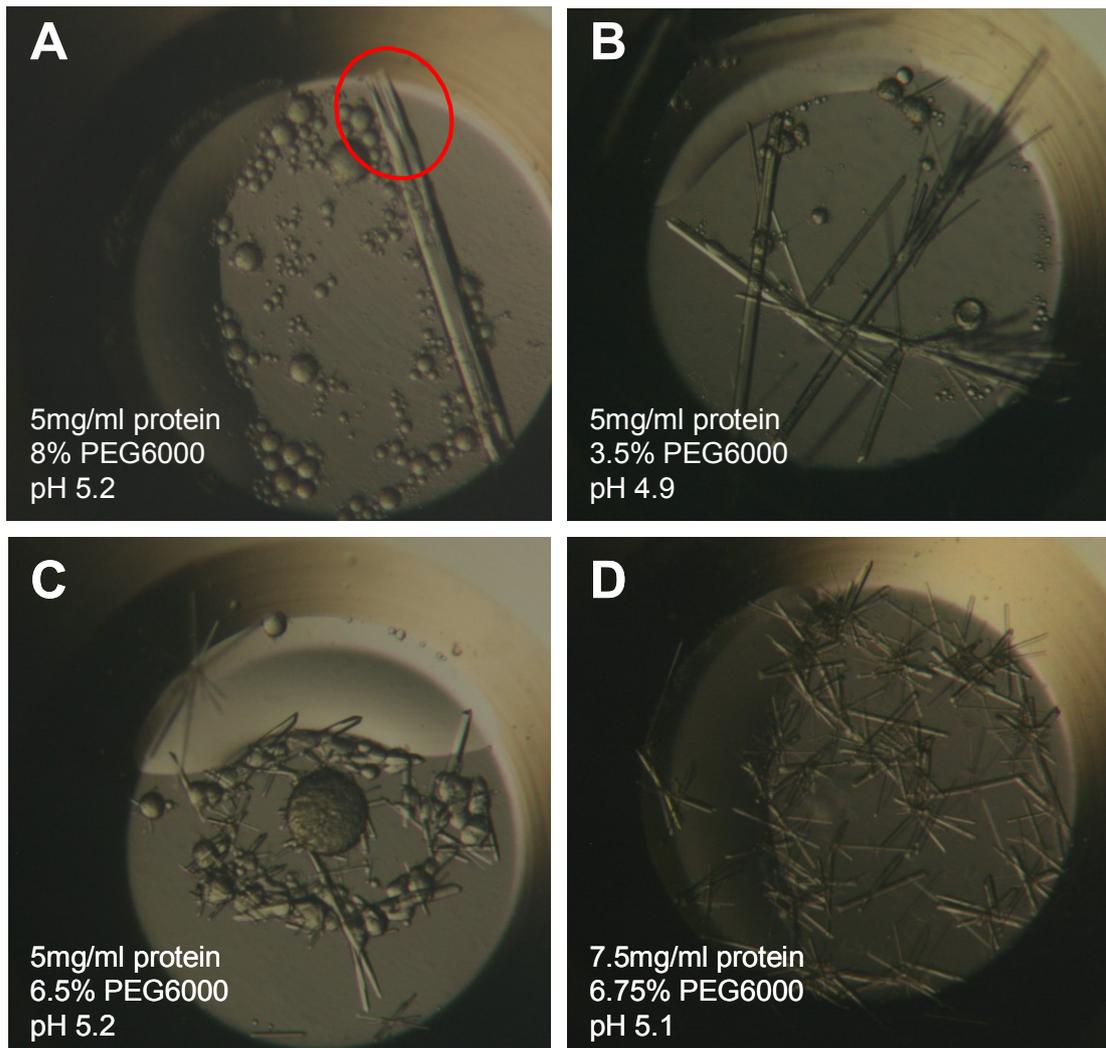
Fig 4.8. Figure showing the variety of crystals produced from optimisation of 'needle like' TmAKR crystal conditions. The red circle highlights poor growth towards the end of the needle. Initial crystallisation conditions for each are shown.

5 X-ray Diffraction and Structure Determination

5.1 Introduction

Using a conventional light microscope it is possible to view details as small as approximately 0.5 μm. This limit is due to the wavelength of the light used since features closer together than approximately one wavelength cannot be resolved. To view finer detail light with a shorter wavelength is required. X-rays with a wavelength of between 0.5 and 1.6Å (0.05-0.16nm) will interact with matter on the correct scale to allow resolution of details on the atomic scale. For X-rays the refractive index of most materials is close to that of a vacuum therefore X-rays cannot be focused like visible light in a conventional light microscope. Being unable to focus the X-rays means the phases of the X-rays cannot be directly observed. This has become known as the phase problem and is discussed in section 5.1.3.

In 1912 it was demonstrated by Max von Laue that crystals of copper sulfate could diffract a beam of X-rays and by 1913 Lawrence Bragg had developed methods to interpret these scattered X-rays. Bragg subsequently solved the first X-ray structure, that of sodium chloride. The structure of sperm whale myoglobin was the first macromolecular structure to be solved, but it was not until the late 1950's that this occurred due to its complexity and computational requirements (Kendrew *et al*., 1958). The structure of sodium chloride was so simple that Bragg was able to derive its structure by reasoning alone and produce the structure straight from the diffraction image. The macromolecular structures solved today still rely on the same basic principles developed by Bragg.

5.1.1 X-ray Sources

To produce X-rays in the laboratory a rotating anode generator may be used. Here electrons are accelerated through a vacuum from a cathode towards a rotating anode (Fig 5.1). When the electrons strike the anode their energy is absorbed causing transitions in the electronic state of the metal atoms. The absorbed energy is then released by the atoms as X-rays of a range of wavelengths. Specific wavelengths of high intensity are produced and can be used for macromolecular structure determination. A rotating anode generator with a copper anode will produce characteristic radiation of 1.54Å alternatively a molybdenum anode may be used to produce X-rays with a wavelength of 0.71Å. The electron beam hitting the anode

creates a large build up of heat and by rotating the anode this heat build up is spread over a larger surface area dissipating the heat. Cold water is also supplied to the inside of the anode to aid with cooling.
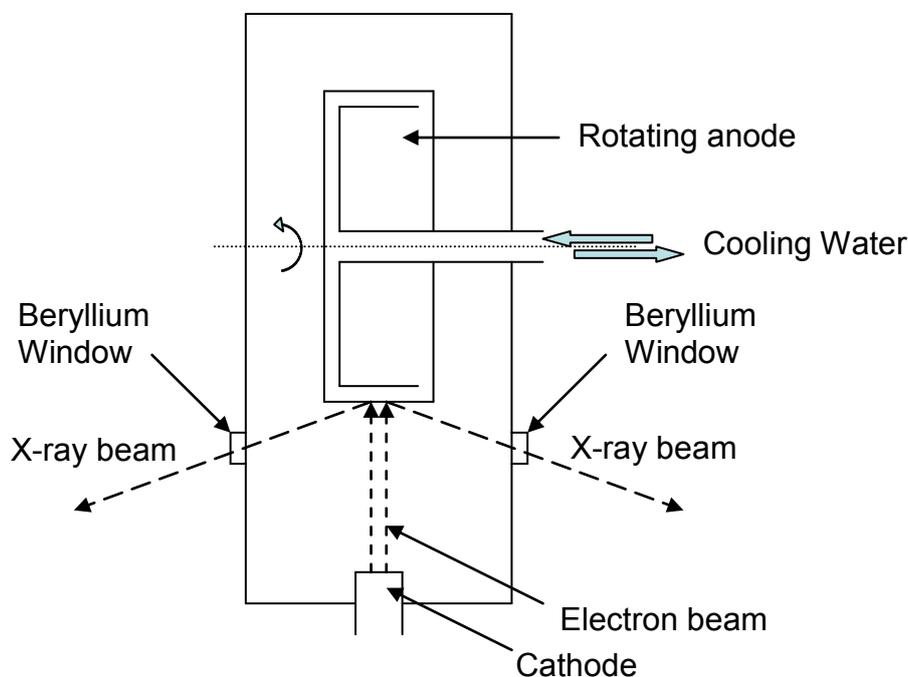
Fig 5.1. Figure showing a schematic representation of a rotating anode X-ray generator. Electrons are produced at the cathode and accelerated towards the anode where X-rays are produced which exit through beryllium windows (transparent to X-rays). The rotating anode is cooled from the inside by a water supply (Adapted from Blow, 2002)

A much more intense source of X-rays is available at synchrotrons, allowing for faster data collection to a higher resolution. At a synchrotron, electrons are generated from a cathode and accelerated by electric fields down a linear accelerator and through a booster ring where they are accelerated further. They finally enter the storage ring where they will reach speeds of over 99% the speed of light. Whilst travelling round the storage ring the electrons paths are affected by a series of large powerful electromagnets which serve to keep the electrons on a circular path. These magnets along with insertion devices know as undulators and wigglers cause the electrons to release energy as electromagnetic radiation. The energy released is over a broad range of wavelengths from X-rays to ultraviolet and visible light. Depending on the insertion device a range of X-rays may be produced, which travel into the

experimental station where a series of mirrors and a monochromator produce a focused beam of a single wavelength.

5.1.2 Diffraction Theory

When a sufficiently ordered crystal is placed in an X-ray beam diffraction effects can be observed. This diffraction is caused by the X-rays being absorbed by electrons and re-emitted in all directions. Since the crystal is a regularly repeating structure these emitted X-rays may reach the detector in phase leading to constructive interference and producing a peak on the detector. The structure of the crystal is related by reciprocal space to the diffraction pattern observed on the detector; the spacing between peaks has an inverse relationship to the spacing of the lattice points of the crystal.

When in 1913 Lawrence Bragg solved the first X-ray structure he reasoned that the repeating structure within the crystal formed as a series of lattice planes which acted like mirrors reflecting the X-rays. The lattice planes can be described by the indices *h*, *k* and *l*. When the spacing between these planes was such that the reflected waves were in phase constructive interference would be observed (Fig 5.2). This became known as Braggs law (Eq. 5.1) and states that two planes will scatter in phase if the difference in path length is equal to a whole number of wavelengths.

$$n\lambda = 2d_{hkl} \sin\theta \qquad \text{Eq 5.1}$$

Where n is an integer, λ is the wavelength, d_{hkl} is the spacing between a particular set of planes *hkl* and θ is the glancing angle of the beam.
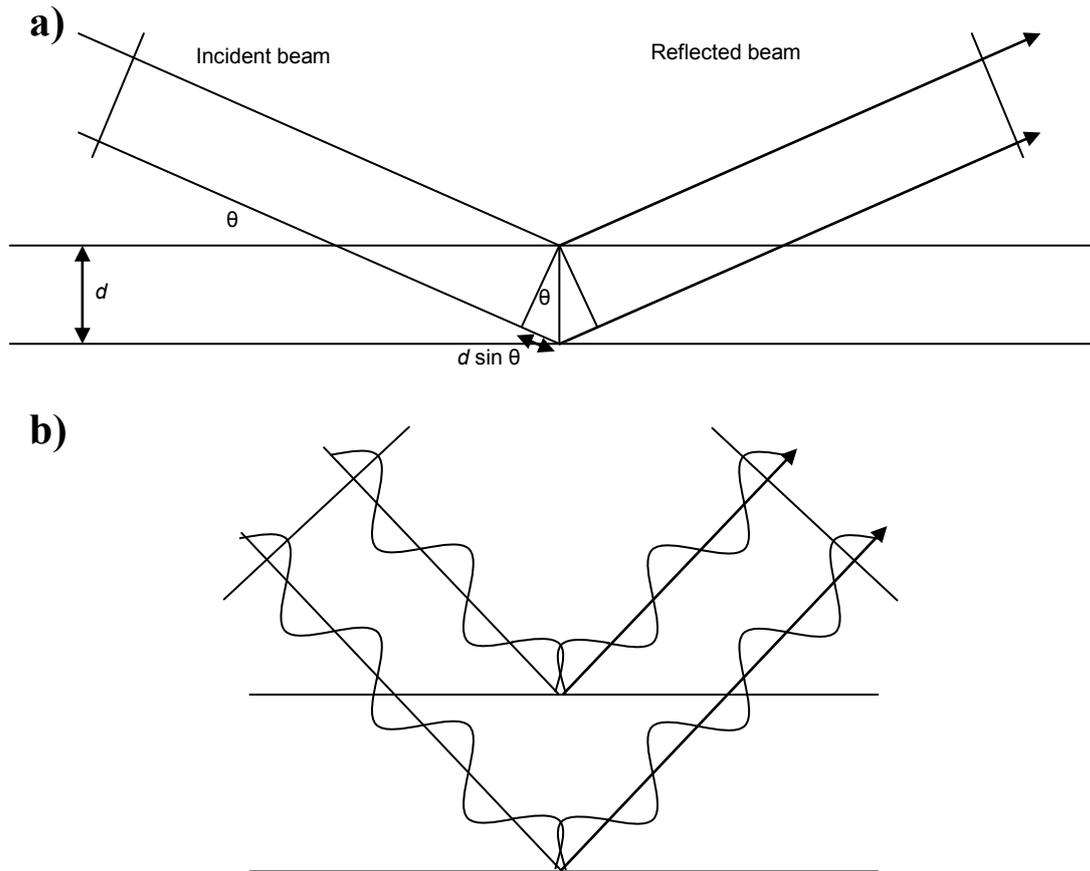
Fig 5.2. Figure showing reflection of incident X-ray beams from a pair of parallel lattice planes. a) When an incident beam is reflected from a lattice plane by the glancing angle θ the beam reflecting off the second plane travels an extra distance equal to $2d\sin\theta$. b) Example where Bragg's law is satisfied; the lower beam travels an extra 2λ and is therefore in phase with the upper beam resulting in constructive interference.

To calculate the electron density for the crystal, first the diffraction data must be converted into structure factors. The structure factor is a description of a wave written as a complex number such that,

$$F = |F|\, e^{i\alpha} \qquad\qquad \text{Eq 5.2}$$

Where F is the structure factor, |F| is the amplitude of the structure factor obtained from experimentally observed intensities and α is its phase. The structure factor F_{hkl} for a particular reflection *hkl* can be expressed as the atomic scattering contribution for each atom within the unit cell.

$$F_{hkl} = \sum_{j=0} f_j \, e^{2\pi i \, (hx_j + ky_j + lz_j)} \qquad\qquad \text{Eq 5.3}$$

Where f_j is the scattering factor of atom j and x_j, y_j and z_j are the fractional coordinates of atom j within the unit cell.

The scattering (electron) density at a point x,y,z can then be calculated from the inverse Fourier transform of the structure factors.

$$\rho(x,y,z) = \frac{1}{V} \sum_{hkl} F_{hkl} \, e^{-2\pi i \, (hx + ky + lz)} \qquad\qquad \text{Eq 5.4}$$

$\rho(x,y,z)$ is the scattering density at a point x,y,z and V the volume of the unit cell. The structure factor is usually expressed as electrons per unit cell where as the density is expressed as electrons per unit volume, hence the term 1/V.

5.1.3 Phase Problem

During a diffraction experiment we are able to measure the intensity of the reflections which can be converted into the amplitudes of the structure factors, however another piece of information is missing, that of the phase, which we cannot measure directly. This missing piece of information has become known as the phase problem, and various different methods exist for solving it.

Two techniques, molecular replacement (MR) and multi-wavelength anomalous dispersion (MAD) have been used here to solve the AKR and Bacterioferritin structures respectively and will be explained in more detail below. In addition to these two techniques multiple isomorphous replacement (MIR) can also be used to generate phases and solve the structure. MIR utilises a native crystal and also a series of crystals either grown in the presence of or soaked in a heavy metal solution. When placed in the X-ray beam the presence of a heavy metal in the unit cell produces differences in the structure factor amplitudes between each derivative and the native crystal. Identification of the heavy metal sites via a Patterson map allows for identification of two potential phases for each derivative, by the use of two or more derivatives the wrong phase can be ruled out and the correct phase identified.

5.1.3.1 Multi-wavelength Anomalous Dispersion (MAD)

Multi-wavelength anomalous dispersion can be used to experimentally derive phases and exploits the ability of heavy atoms to absorb X-rays at specific wavelengths. Any particular atom can absorb energy and promote an electron from one orbital to a higher orbital, and subsequently release this energy again as the electron returns to its normal transition state. The electrons in certain orbitals of heavy elements (Fe and above) are able to absorb X-rays of wavelengths convenient for diffraction experiments and produces an effect called anomalous scattering. We are able to experimentally derive phases from this effect since anomalous dispersion creates a violation of Friedel's law which states that a pair of symmetrical reflections *hkl* has the same amplitude and opposite phase as the corresponding reflection *-h-k-l*. When anomalous dispersion occurs this law is broken (Fig 5.3) and *-h-k-l* has a different intensity to the reflection *hkl*, this pair of reflections are known as a bijvoet pair and the difference in their amplitudes (ΔF) the bijvoet difference.
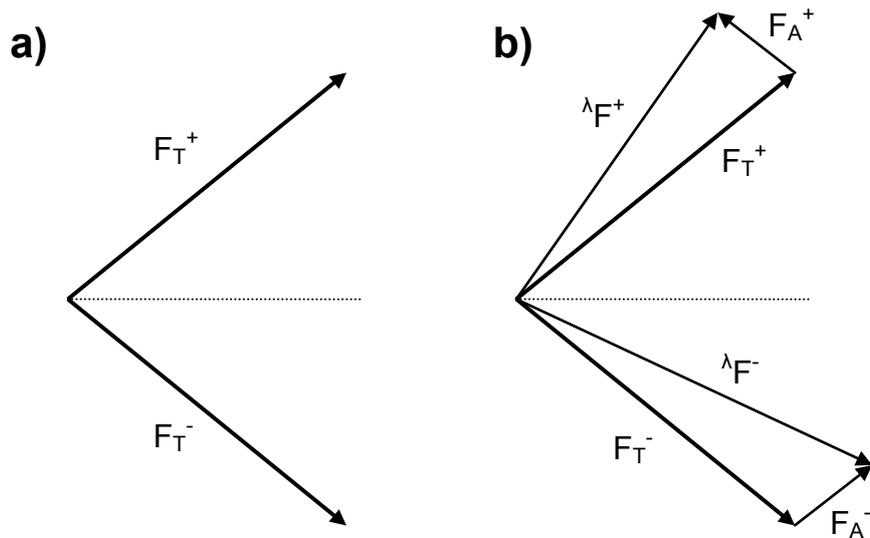
Fig 5.3. Figure demonstrating Friedel's law and its disruption by anomalous scattering effects. a) Normal components of reflections hkl (F_T^+) and –h-k-l (F_T^-) obeying Friedel's law, amplitudes are equal and phases are opposite. b) The anomalous component F_A contributes to the normal scattering resulting in the total observed scattering $^\lambda F$ no longer obeying Friedel's law.

The total scattering is a result of the normal scattering and anomalous scattering and can be expressed as

$$f(\lambda) = f_0 + f'(\lambda) + if''(\lambda) \qquad \text{Eq 5.5}$$

Where f_0 is the normal atomic scattering factor, f' is the dispersive term and real part of the anomalous scattering component and f'' is the imaginary component of anomalous scattering. The values of f' and f'' vary with the wavelength (Fig 5.4) with f'' having a maximum value at the absorption edge and f' being at its minimum at the inflection point of the absorption peak.
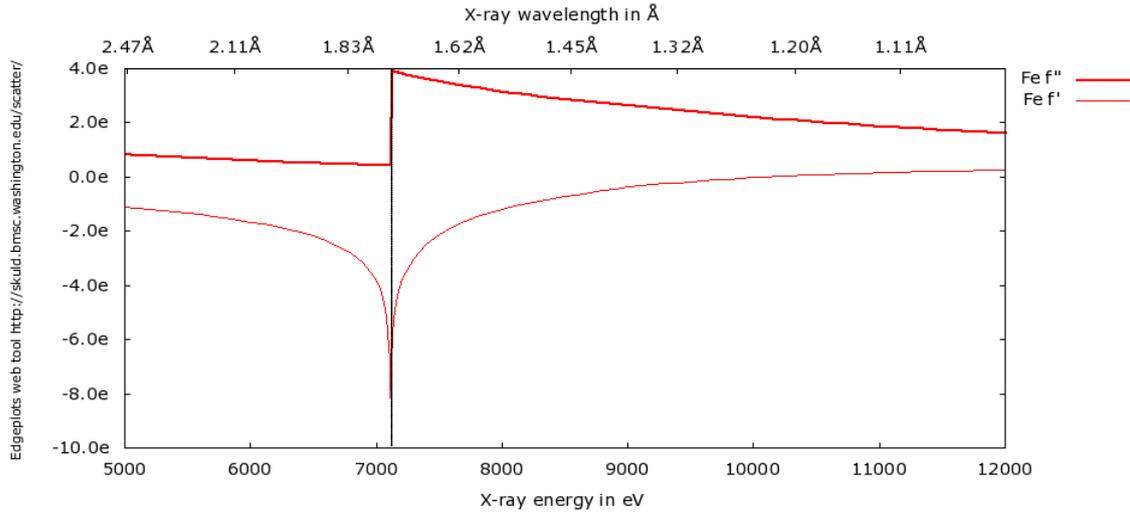
Fig 5.4. Theoretical X-ray absorption curve for the Fe K edge, showing change in anomalous scattering components f' and f''. Figure produced from http://skuld.bmsc.washington.edu/scatter/AS_form.html

In the 1980s Hendrickson was able to use anomalous dispersion to solve to a protein structure (Hendrickson, 1985) by using the equations of Karle (1980). Accounting for the anomalous difference the structure factor equation becomes,

$$^{\lambda}F(h) = {}^{0}F_{T}(h) + \sum_{k} {}^{0}F_{Ak}(h) \left[(f'_{k}/f^{0}_{k}) + (if''_{k}/f^{0}_{k}) \right] \qquad \text{Eq 5.6}$$

Where $^{\lambda}F(h)$ is the structure factor at wavelength λ, $^{0}F_{T}(h)$ is the structure factor from all normal scatterers and $^{0}F_{Ak}(h)$ is the normal contribution from all anomalous scatterers. The final part of the equation defines the amount of anomalous scattering dependant upon the values of f' and f'' which are dependant upon the wavelength. By collecting datasets at multiple wavelengths and therefore differing contributions of anomalous scattering components we are able to solve the equation to provide us with values for F_{T}, F_{A} and $\Delta\varphi$, the difference in phase angle between normal and anomalous components. At least two different wavelengths are required to solve the equations but it is preferable to have three or four to provide a more reliable solution. These datasets are usually collected at wavelengths corresponding to the maximum value of f'' (maximum absorption), the minimum value of f' (minimum normal scattering), and one or two remote data sets where f' is close to its normal value.

Heavy atom positions may be identified from an anomalous difference Patterson map using the square of the Bijvoet differences. This map contains peaks at the vectors corresponding to the heavy atom positions and can be used to give estimates for heavy atom positions within the unit cell. Once positions of the anomalous scatterers are known their phases, φ_A, can be calculated and used to calculate phases for the whole structure, φ_T, since $\varphi_T = \varphi_A + \Delta\varphi$.

5.1.3.2 Molecular Replacement (MR)

Experimental phasing and the collection of multiple datasets is not always necessary. If the protein of interest is sufficiently similar to another protein which has a structure available the situation arises where it may be possible to use the phases of the known structure as initial phases for the unknown structure. Solving the phase problem in this way is known as molecular replacement. Molecular replacement requires the model structure to be sufficiently similar to the unknown structure, usually sharing at least 25% amino acid identity. In performing a molecular replacement we need to initially calculate 6 parameters, 3 rotation angles and 3 translation vectors. These 6 parameters describe how the known molecular structure relates to the unknown. To speed up the computation of these parameters the calculations are separated into two stages (two 3 dimensional searches, rather than one 6 dimensional search). The first stage concerns describing the 3 rotation angles, this is done by exploiting the Patterson map which does not depend upon any choice of origin for the molecule or require any phase information. The origin of the Patterson is where vectors between atoms are zero and allows the rotation to be searched for, independent of any translation which may exist. By taking the Patterson of the known molecule and rotating it over the Patterson of the unknown structure we can identify where the two Pattersons agree with each other and therefore the rotation angle between the two molecules. The second stage is to find where the origin of the model should be placed in the unit cell of the unknown structure to derive a diffraction pattern similar to that of the unknown structure. The method is to place the model at all points in the unit cell and calculate the corresponding Patterson function and compare that to the actual Patterson from the unknown structure.

5.1.4 Model building and refinement

Once the phase problem has been solved and electron density calculated, the model can be improved by manually altering the positions of the atoms to better fit the density. Some features may be observed in the electron density map and allow positioning of atoms. Also the use of a 'difference' (F_0-F_C) map can give an indication as to atom positions. The F_0-F_C map represents the difference between the observed (F_0) and calculated (F_C) structure factors and shows areas where the model disagrees with the observed experimental data.

After the electron density has been interpreted and the model altered to take this into account, structural refinement may take place. Structural refinement is carried out computationally to alter the model to give the best fit to the observed structure factors. To refine the position of every atom individually would require four variables (x, y, z, B-factor) per atom and become computationally expensive, instead by placing restraints on the positions of atoms this can be reduced to an average of four per residue. The peptide bond is planar allowing restraints to be imposed to retain this geometry and in doing so allows the main chain of each residue to be defined by only two variables, the φ and ψ angles around the Cα atom. Each side chain can also be represented by a series of dihedral angles, χ_1, χ_2 etc… dependant on the length of the side chain. At the end of the structural refinement a new map is calculated for the model, interpreted and the model altered accordingly. This cyclical procedure is repeated until no further appreciable improvement is seen.

Throughout the refinement procedure the quality of the model is followed by monitoring the R factor which is used as a measure of the fit of the model to the observed data.

$$R = \frac{\sum_h ||F_{obs}| - |F_{calc}||}{\sum_h |F_{obs}|} \qquad \text{Eq 5.7}$$

Where F_{obs} is the observed structure factors and F_{calc} is the calculated structure factors from the model over set of reflections h. Throughout refinement the R factor decreases as the model improves, with a value of zero representing perfect agreement between the observed intensities and the model. The R factor can provide an indication as to when refinement has finished, by monitoring it's progress and observing when its value no longer improves. The R factor suffers from bias

introduced by using the same data for refinement as well as for monitoring. The Free R factor (R_{free}) can be used to overcome this bias and takes a set of the diffraction data (usually 5%) which is not used for the refinement procedure and treats it in the same way as for the R factor. R_{free} is typically larger than the R factor but is a more reliable estimate of model quality.

5.1.4 Validation

The R factor and R_{free} are both commonly used to assess model quality throughout the refinement procedure. When the R factors do not improve any further then refinement is normally terminated and a final model produced. To further assess the quality of this final model several programs such as PROCHECK (Laskowski *et al*., 1993) and SFCHECK (Vaguine *et al*., 1999) may be used. These programs analyse the final structure to check parameters such as bond angles, bond lengths, stereochemistry and planarity of peptide bonds and cyclic side chains. These values are compared to expected values for the appropriate resolution.

A Ramachandran plot is usually also produced which plots the phi and psi bond angles for the main chain. Due to steric hindrance the main chain may only adopt certain angles and residues should not appear in the disallowed regions of the plot, and should be seen in certain regions relating to different types of secondary structure.

5.2 Materials and Methods

5.2.1 TmAKR

5.2.1.1 Cryoprotectants

Single crystals of TmAKR were harvested from the crystallisation drop and placed into appropriate mother liquor (Table 5.1) for 1 minute prior to flash freezing in liquid nitrogen. Crystals were stored under liquid nitrogen until required.

Crystal	Si18	SCW1	SCW10
Glycerol %(v/v)	30	30	30
Ammonium sulfate %(Sat.)	20	30	30
NADPH (Molar eq.)	0	1.2	1.2
Acetic Acid (mM)	0	50	50
Citric Acid (mM)	50	0	0
Buffer pH	5.1	4.5	4.5

Table 5.1. Composition of mother liquors used for soaking crystals of TmAKR prior to flash freezing.

5.2.1.2 Data collection

5.2.1.2.1 'Apo' TmAKR

Initial data were collected at the SRS at Daresbury with the crystal 'Si18', cryo-cooled at 100K. This crystal was grown in the absence of NADPH and represents the apo form of the enzyme. The crystal diffracted to 2.64Å and data were collected with an oscillation range of 0.5°, crystal to detector distance of 247 mm and slit widths of 0.2x0.2 mm.

5.2.1.2.2 'Holo' TmAKR

Data were collected from crystals co-crystallised in the presence of NADPH cofactor. All data for crystals co-crystallised with NADPH were collected at the SRS at Daresbury using single cryo-cooled crystals. Data set SCW10 was collected with smaller slit widths (0.1x0.1 mm) than previous data sets to reduce spot size and number of overlaps, improving measurements for this data set. Data collection parameters for holo crystals are summarised in table 5.2.

Crystal	SCW1	SCW10
X-ray wavelenth (Å)	1.1200	1.2200
Oscillation (°)	0.25	0.25
Detector distance (mm)	180	180.08
Slit widths (mm)	0.2	0.1
Max resolution (Å)	2.2	2.4

Table 5.2. Data collection parameters for 'holo' crystals of TmAKR.

5.2.1.3 Data Processing

All data sets were processed using DENZO and SCALEPACK (Otwinowski, 1997). 'Si18' was indexed in space group P422 and molecular replacement performed by MOLREP (Vagin, 1997) using 1PZ0 (AKR11A1 from B. subtilis; 34% sequence identity) as a search model. Two monomers were found in the asymmetric unit and the model sequence altered manually according to an amino acid sequence alignment of TmAKR and 1PZ0 (Bioedit; Hall, 1999; Appendix 5). This model was used for further model building and refinement using COOT (Emsley and Cowtan, 2004) and REFMAC (Murshudov, 1997), to produce an intermediate model.

Data set SCW1 was processed in space group P422 and the intermediate model of Si18 was used as a search model for molecular replacement using MOLREP Which was unsuccessful. Data were submitted to the Zanuda server (http://www.ysbl.york.ac.uk/) to correct space group assignment. SCW10 data were processed in P222 and molecular replacement performed using MOLREP and the intermediate Si18 model as a search model as for data set SCW1, and was also unsuccessful. The program REINDEX (CCP4, 1994) was used with the operator h, k, l/3 to reduce the unit cell of SCW10. Molecular replacement was performed as before using the intermediate Si18 model with the reindexed SCW10 data. Two monomers were found in the asymmetric unit, and model building and refinement was performed with COOT and REFMAC.

The final model of SCW10 was used as a search model for molecular replacement with the Si18 data. Model building and refinement were performed with COOT and REFMAC to produce the final TmAKR model.

5.2.2 Bacterioferritin

Single crystals of bacterioferritin were harvested from the crystallisation drop and placed into a mother liquor of 30% (v/v) glycerol, 60% (Sat.) ammonium sulfate, 20mM Tris-HCl pH 7.5 for 1 minute prior to flash freezing in liquid nitrogen. Crystals were stored under liquid nitrogen until required.

X-ray data were collected at the SRS at Daresbury, using a single cryo-cooled crystal at 100K and a MAR225 CCD detector. An EXAFS (Extended X-ray Absorption Fine Structure) scan at the Fe K edge was performed using a C-TRAIN Fluorescence detector to confirm the presence of iron as suggested by the red colour of the crystals and the peak absorption was found to correspond to a wavelength of 1.739Å. Data were collected at a peak wavelength of 1.739Å and a remote wavelength of 1.729Å. Both data sets were collected with 30 second exposures using an oscillation angle of 0.3°, crystal to detector distance of 90 mm and slit widths of 0.1x0.1 mm. Data were processed using DENZO and SCALEPACK (Otwinowski, 1997). The program TRUNCATE (French, 1978) was used to calculate structure factors and MLPHARE (CCP4, 1994) for phase calculation. Averaging and phase combination was performed with DM (Cowtan, 1994).

Iron atom positions were identified using SHELXD of the SHELX package (Sheldrick, 2008) and used to create MAD phases. A partial model made up of polyalanine α-helices was manually build into the density using COOT (Emsley and Cowtan, 2004); phased refinement of partial model phases and combination with anomalous phases was performed using the program REFMAC (Murshudov, 1997). Several rounds of model building and refinement were carried out to produce a model containing four helices. Sidechains were then identified and built into the model. The putative amino acid sequence determined from the electron density was then input into BLAST (Altschul, 1990) to allow identification of the protein as *E. coli* bacterioferritin.

A dimer of *E. coli* bacterioferritin (PDB: 1BFR) was used as a model for molecular replacement using MOLREP (Vagin, 1997). Further dimers, to produce a total of 12 monomers, were found neighbouring the first using molecular replacement and the symmetry operators from the iron positions using LSQKAB (Kabsch, 1976). This model was used for further refinement using COOT and REFMAC, and ordered water molecules were added using COOT and ARP/wARP (Perrakis, 1997).

5.2.3 Proton Induced X-ray Emission (μPIXE)

μPIXE measurements were performed at the Ion Beam Centre at the University of Surrey. A 2.0 MeV 1.5 μm diameter proton beam was used to induce X-ray emission from 0.2 μL liquid protein samples dried onto 2 μm thick mylar film in a vacuum. X-rays were detected by a high energy resolution solid-state lithium-drifted silicon detector. The proton beam was scanned across the drop in X and Y, and by defining software windows round the X-ray peaks, the signal from elements of interest were sorted into elemental maps. In order to try to quantify the element concentrations, point spectra were collected at four selected points in the sample and also at points on the buffer and the backing foil. The spectra were analysed by the DAN32 (Grime, 2004) and GUPIX software (Campbell *et al*., 2000).

The μPIXE method has been used successfully in both identifying elements in proteins and in measuring their stoichiometric ratio to an accuracy of between 10 and 20% on over 70 samples (Garman and Grime, 2005). Using the sulfur peak to give an internal normalisation of the sulfur atoms from the known cysteines and methionines, the method is calibrated per protein molecule. This work was performed in collaboration with Elspeth Garman (University of Oxford)

5.3 Results and Discussion

5.3.1 TmAKR

5.3.1.1 Data Collection

For the crystal 'Si18' diffraction was observed to 2.64Å and 174 frames of data collected. The data were indexed in space group P422 with the unit cell parameters a = 141.1Å, b = 141.1Å, c = 84.1Å, $\alpha = \beta = \gamma = 90.0°$.

Data sets were also collected from crystals formed in the presence of NADPH. The crystal 'SCW1' diffracted to 2.2Å and 174 frames of data were collected. The data were indexed in space group P422 with the unit cell parameters a = 141.6Å, b = 141.6Å, c = 169.5Å, $\alpha = \beta = \gamma = 90.0°$. This crystal was indexed in the same space group as the apo form with a doubling of unit cell parameter c (84.1Å – 169.5Å). Data collection for this crystal was not completed due to a synchrotron failure and has a final completeness of 82%. A further holo data set was collected from crystal 'SCW10' which diffracted to 2.4Å and 421 frames of data were collected. The data were indexed in space group P222 with the unit cell parameters a = 140.9Å, b = 140.9Å, c = 252.7Å, $\alpha = \beta = \gamma = 90.0°$. This crystal was indexed in a lower symmetry space group than the former two, and with a tripling of the unit cell parameter in b (84.1Å – 252.7Å) relative to Si18. Data collection statistics for all three data sets are shown in table 5.3. Representative X-ray diffraction frame for dataset Si18 is shown in figure 5.5, and for dataset SCW10 in figure 5.6.

Crystal	Si18	SCW1	SCW10
Space group	P4 2_1 2	P4 2_1 2	P2$_1$ 2$_1$ 2
Unit cell lengths (a,b,c) Å	141.1, 141.1, 84.1	141.6, 141.6, 169.5	140.9, 104.9, 252.7
Unit cell angles (α,β,γ) °	90, 90, 90	90, 90, 90	90, 90, 90
Resolution (Å)	2.64	2.24	2.77
Completeness (%)	99.5 (97.2)	82.9 (85.5)	87.6 (86.2)
Redundancy	6.0	3.0	2.4
No. of reflections	142327	189463	393453
No. of unique reflections	23900	59252	167009
Wilson B-factor ($Å^2$)	61.8	63.2	63.1
R_{sym} (%)	8.5 (50.8)	9.3 (57.7)	12.9 (57.5)
$<I>/<\sigma I>$	20.9 (2.90)	10.4 (1.7)	7.67 (1.64)
I>3σI (%)	77.5 (36.1)	49.2 (12.9)	44.6 (12.4)

Table 5.3. Table showing the data collection statistics for three TmAKR data sets. Values in parenthesis are for highest resolution shell. $R_{sym} = \sum|I - <I>| / \sum(I)$ where I is the intensity of a reflection.
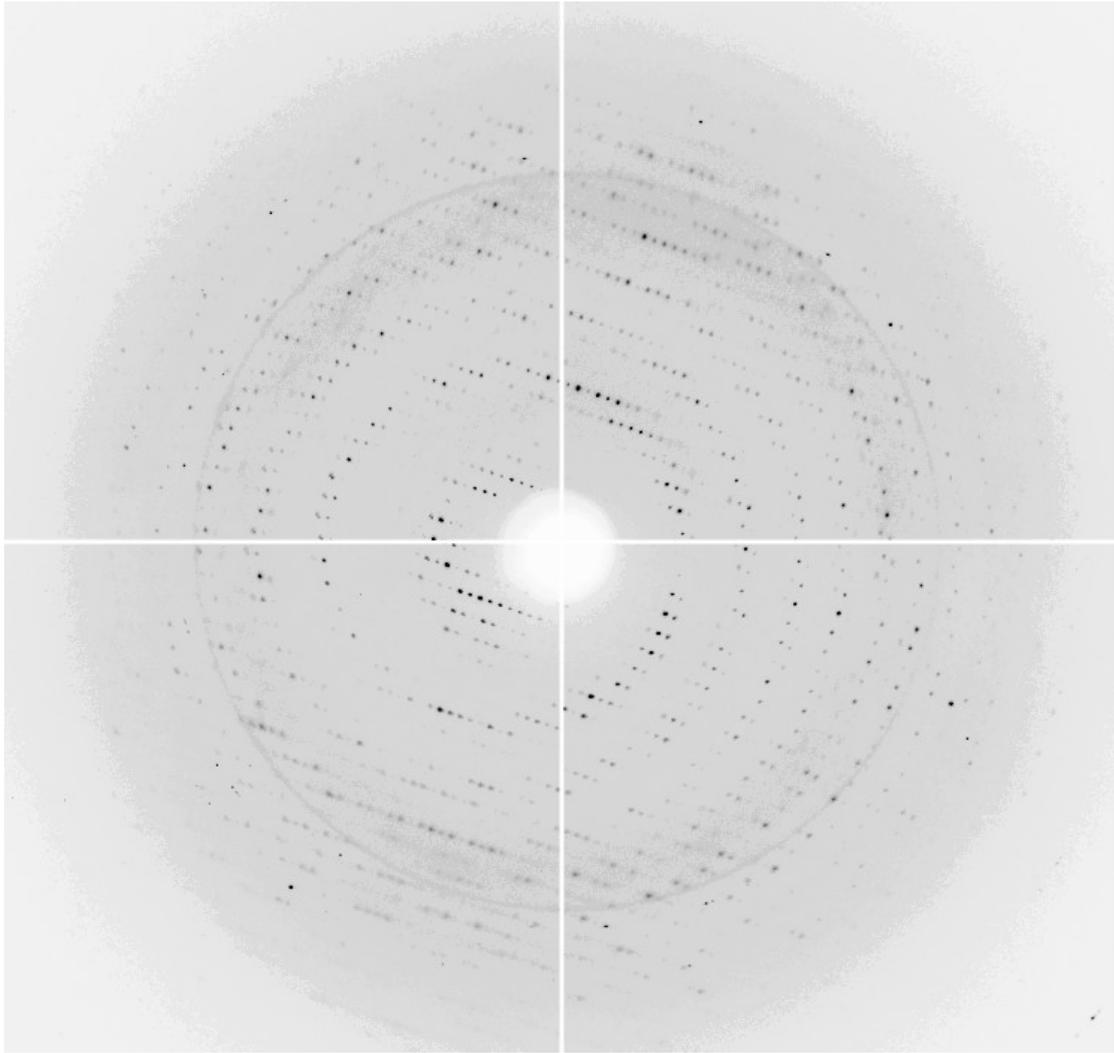
Fig 5.5. Figure showing a single representative X-ray diffraction image for dataset Si18 collected at Daresbury SRS, showing diffraction to 2.64Å resolution.
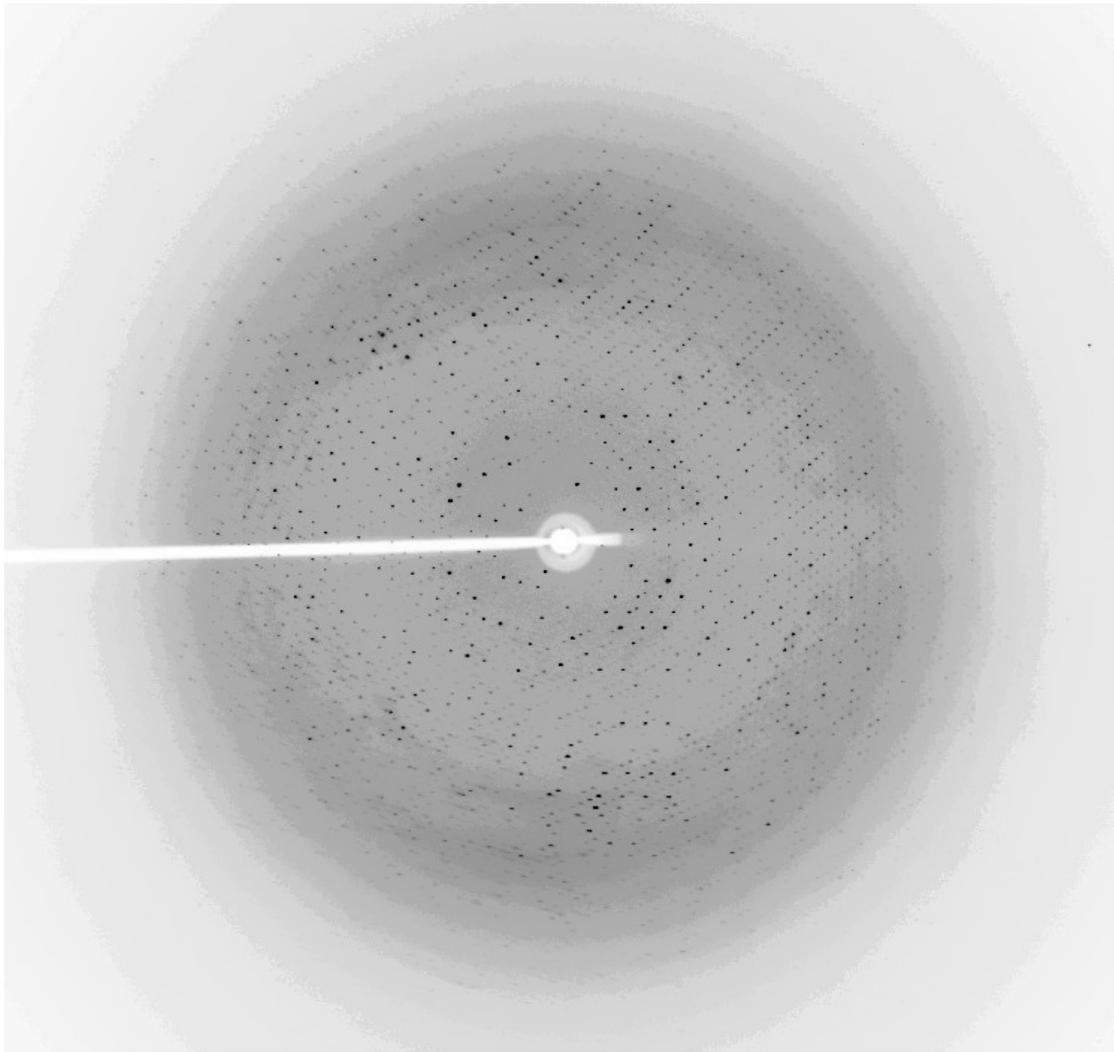
Fig 5.6. Figure showing a single Representative X-ray diffraction image for dataset SCW10 collected at Daresbury SRS, showing diffraction to 2.4Å resolution.

5.3.1.2 Structure Solution and Refinement

Molecular replacement was performed for the Si18 data set by MOLREP using the AKR11A1 structure from *B. subtilis* (PDB: 1PZ0) as a search model, and two monomers were found in the asymmetric unit. The amino acid sequence alignment (Appendix 5) for TmAKR and 1PZ0, was used to manually alter the model sequence to match that of the TmAKR. This model was used for further model building and refinement using COOT and REFMAC. The intermediate model at this stage (R = 0.263; R_{free} = 0.321) consisted of 276 amino acids (333 sequence total) with two loops (27-32 and 212-245) and the C-terminus (316-333) not appearing in the density which are presumed to be disordered.

The SCW1 data set was input into MOLREP using the intermediate model of Si18 as a search model. MOLREP detected pseudotranslation at 0.5c + 0.1a. No correct solution was found which would refine well in REFMAC. Despite the failure of molecular replacement it was still desirable to know if the co-factor, NADPH, was present in the structure. Therefore the data were submitted to the Zanuda server which suggested the space group to be $P2_12_12_1$ and implies perfect twinning by hemihedry (personal communication, A. Lebedev).

Molecular replacement for the SCW10 data were performed using MOLREP and the intermediate Si18 model as a search model. MOLREP also detected pseudotranslation in the data (0.33c; 0.66c), and initial molecular replacement was not successful. The data were submitted to the Zanuda server , and the space group $P42_12$ identified as the most likely space group. In contrast to crystal SCW1, SCW10 was not twinned, and the program REINDEX (CCP4, 1994) was used with the operator (h,k,l/3) to reduce the unit cell according to pseudotranslation to similar dimensions to that of the apo form. Molecular replacement was performed as before using the Si18 model with the reindexed SCW10 data. Two monomers were found in the asymmetric unit with an R-factor of 0.255 (R_{free} = 0.298). Model building and refinement was performed with COOT and REFMAC, and the missing loops were extended according to the electron density. An additional piece of density was found in each subunit, and a short polyalanine chain build into the density. The amino acids in these extra chains were assigned on the basis of the electron density present, and by comparison with the amino acid sequence are thought to be from one of the missing loops (225-230).

After model building for SCW10 was complete, final R = 0.211 (R_{free} = 0.277), this final SCW10 model was used as a search model for molecular replacement with the Si18 data in an attempt to extend loops missing from the previous intermediate model. In the final Si18 model, subunit A contained 287 amino acids and subunit B contained 292 amino acids (333 total amino acids per subunit). In subunit A three loops (27-32, 216-224 and 231-243) and the C-terminus (316-333) were not found in the density, and in subunit B three loops (29-32, 216-223 and 230-241) and the C-terminus (318-333) were also not found in the density. Final refinement statistics for 'Si18' are presented in table 5.4.

R value (%)	21.1
Free R value (%)	27.7
Mean B value ($Å^2$)	49.1
Total No. of atoms	4695
No. observations	23048
No. water atoms	124
Average B-factor (Protein: $Å^2$)	48.6
Average B-factor (Solvent: $Å^2$)	61.8
RMS deviations from ideality	
Bond lengths (Å)	0.012 (0.022)
Bond angles (deg.)	1.557 (1.997)
Planar groups (Å)	0.007 (0.021)

Table 5.4. Refinement statistics for final model of TmAKR. Values in parenthesis are target values.

5.3.1.3 Data Validation

The final TmAKR structure was analysed with the programs PROCHECK and SFCHECK, to validate bond lengths, bond angles, planar groups, chirality and overall stereochemical quality of the model. The final R-factor for the TmAKR model was 21.3% (R_{free} = 27.7%), in the resolution range 28.7 - 2.70Å. The Cruickshank DPI (Diffraction-component Precision Index) was 0.340Å for the coordinate error of an atom with average B-factor. The stereochemical quality of the model proved to be good with a G-factor of -0.2, which is better than expected (-0.7) for the resolution. All main chain and side chain parameters were found to be within or better than the expected range for the resolution (Appendix 6).

A Ramachandran plot (Fig 5.7) was produced using the program PROCHECK to assess the quality of the model with respect to the phi-psi bond angles of the main chain backbone. The plot showed that out of 481 (non-pro/gly) amino acids, 88.6% of these were present in the most favoured regions and 10.2% in the additionally allowed regions. Five residues (1%) were found in the generously allowed regions and one residue (0.2%) found in the disallowed regions. 88.1% of residues in the favoured regions is higher than expected (72.9%) for the resolution range.
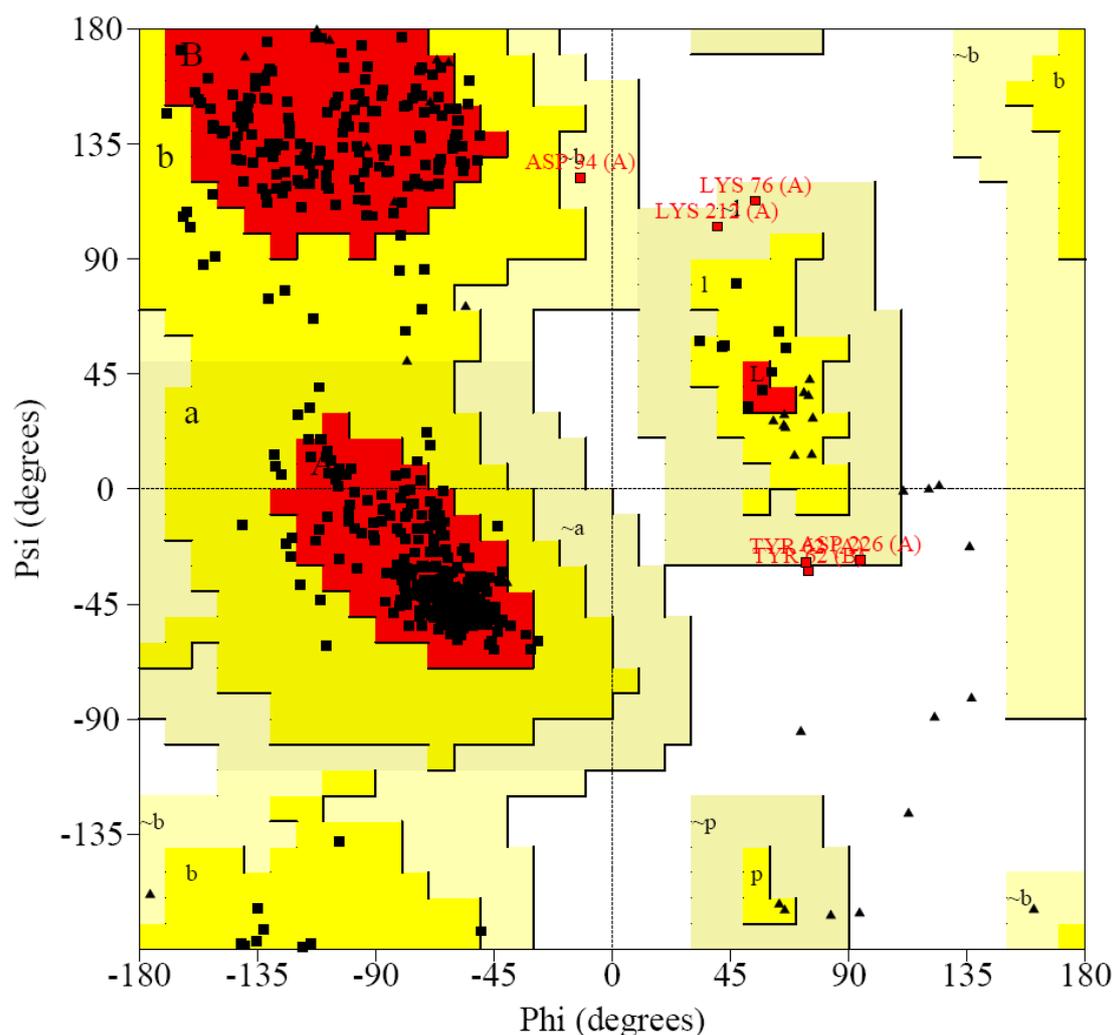
Plot statistics

Residues in most favoured regions [A,B,L]	426	88.6%
Residues in additional allowed regions [a,b,l,p]	49	10.2%
Residues in generously allowed regions [~a,~b,~l,~p]	5	1.0%
Residues in disallowed regions	1	0.2%
	----	------
Number of non-glycine and non-proline residues	481	100.0%
Number of end-residues (excl. Gly and Pro)	119	
Number of glycine residues (shown as triangles)	50	
Number of proline residues	36	

Total number of residues	686	

Fig 5.7. Ramachandran plot for TmAKR dimer, showing 88.6% of all residues in most favoured regions. Those residues in generously allowed regions and disallowed regions are labelled. Figure produced using PROCHECK.

5.3.2 Bacterioferritin

5.3.2.1 Structure Solution and Refinement

The Fluorescence scan at the Fe K edge (Fig 5.5) showed the presence of iron and identified the wavelength to collect the 'peak' data set, where f'' has its maximum value. Both peak and remote data sets were collected to a maximum resolution of 1.9Å (Fig 5.6) and were indexed using DENZO in the space group $P4_22_12$, with unit cell parameters a=208.1 b=208.1 c=142.8Å and α=β=γ=90.00°. Processing statistics are presented in table 5.5. Anomalous phasing with MLPHARE provided initial phases for the structure with a figure of merit of 0.136 to 1.9Å (0.27 to 3.5Å) allowing identification of 12 iron sites within the asymmetric unit. Six fold averaging was applied to improve the phases further and allow identification of helices and the backbone to be traced. Helices were built into the electron density one by one, model phases were combined with anomalous phases and averaged, and new electron density maps calculated after addition of each segment of α-helix. After addition of 4 helices it was possible to build loops connecting the helices and putatively identify a number of side chains. This putative amino acid sequence identified from the electron density was used as a search sequence with BLAST. *E. coli* bacterioferritin was found have the highest similarity to the putative input sequence.

Data set	Remote	Peak
Wavelength (Å)	1.729	1.739
Resolution range (Å)	25 - 1.9 (1.93-1.90)	25 - 1.9 (1.93-1.90)
Completeness (%)	99.2 (90.34)	97.7 (78.9)
No. of reflections	1741445	1763956
No. of unique reflections*	470876	460546
Wilson B value (A^3)	15.4	18.8
$<I>/<\sigma I>$	10.7 (1.02)	11.0 (2.14)
$I>3\sigma I$ (%)	48.9 (11.3)	60.8 (25.2)
R_{sym} (%)	12.8 (78.9)	10.0 (37.7)

Table 5.5. Processing statistics for peak and remote data sets of *E. coli* bacterioferritin. $R_{sym}=\sum_h\sum_J|<I_h>-I_J(h)|/\sum_h\sum_JI(h)$, where I(h) is the intensity of reflection h. $\sum_h$ is the sum over all reflections and $\sum_J$ is the sum over J measurements of the reflection. *Anomalous equivalents were scaled separately. Values in parentheses are for the highest resolution shell.
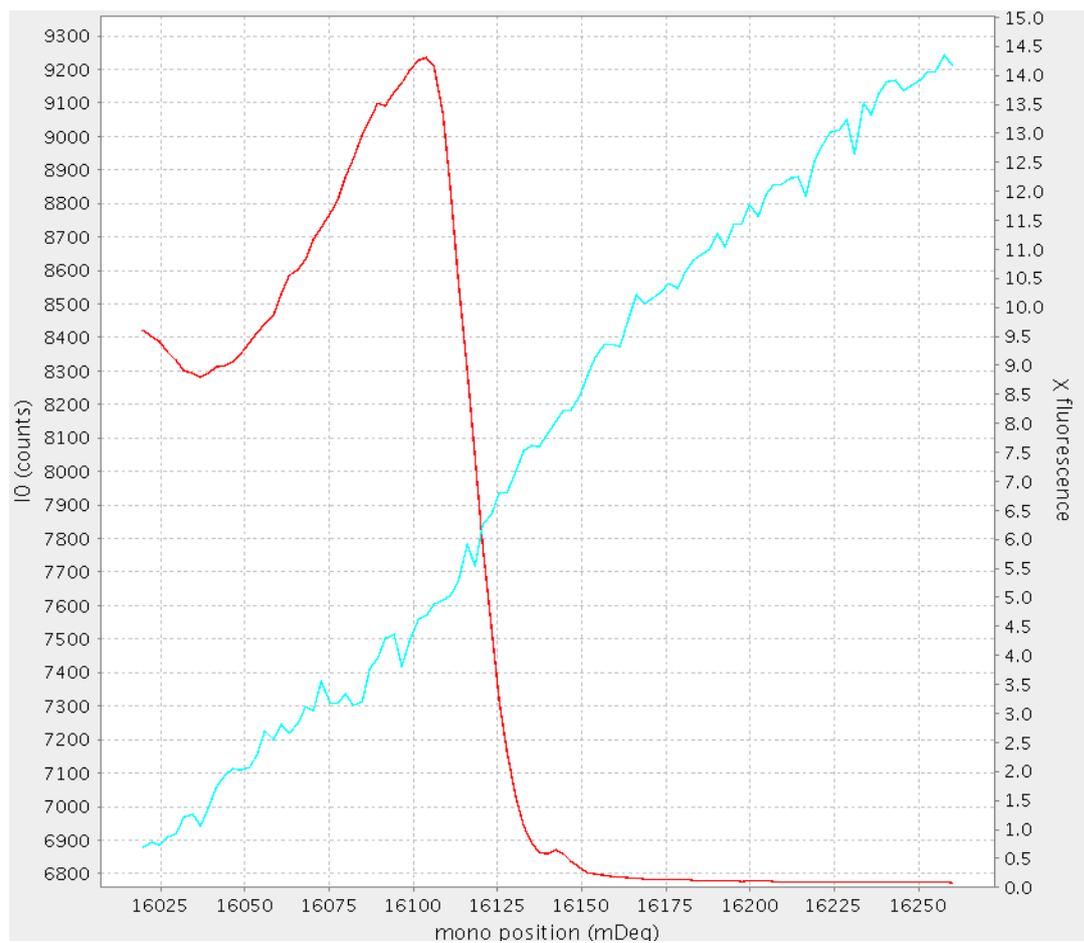
Fig 5.8. EXAFS scan of the Fe K edge used to identify the wavelengths used for data collection for *E. coli* bacterioferritin. Red line shows X-ray fluorescence.
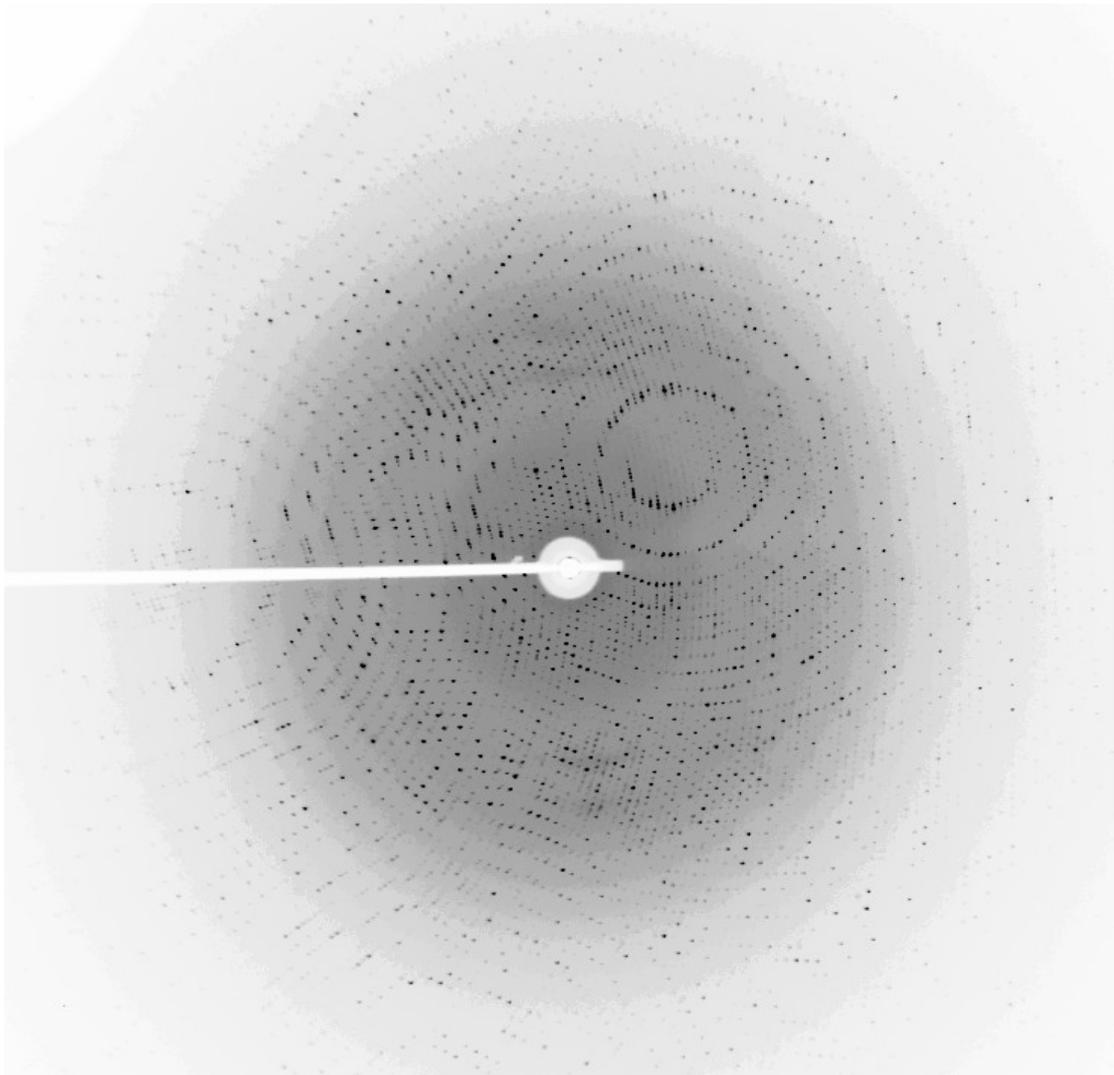
Fig 5.9. Figure showing a representative X-ray diffraction image for *E. coli* bacterioferritin peak data set collected at Daresbury SRS, showing diffraction to 1.9Å resolution.

Since a structure of *E. coli* bacterioferritin already existed (PDB: 1BFR), this was used as a model for molecular replacement. Twelve monomers were built into the density by MOLREP. Each monomer at this stage contained two manganese atoms as found in 1BFR, and a heme molecule located at the two fold axis between two monomers. Side chain positions were altered where appropriate and ordered water molecules added and their positions manually checked. Manganese atoms were replaced with zinc as suggested by μPIXE and analysis of the X-ray data (Section 5.3.2.2). Refinement statistics are presented in table 5.6.

R value	0.181
Free R value	0.217
Mean B value Å^2	18.9
Total No. of atoms	18819
No. water atoms	2209
No. Iron atoms	6
No. Zinc atoms	24
Average B-factor (Protein: Å^2)	18.0
Average B-factor (Solvent: Å^2)	26.9
Average B-factor (Heme: Å^2)	23.2
Average B-factor (Iron: Å^2)	21.5
Average B-factor (Zinc1: Å^2)	21.3
Average B-factor (Zinc2: Å^2)	24.2
RMS deviations from ideality	
Bond lengths (Å)	0.020 (0.021)
Bond angles (deg.)	1.704 (2.071)
Planar groups (Å)	0.007 (0.020)

Table 5.6. Refinement statistics for *E. coli* bacterioferritin. Values in parenthesis are target values. $R_{cryst} = \sum||Fo|-|Fc||/\sum|Fo|$

5.3.2.2 Proton Induced X-ray Emission (µPIXE)

The presence of iron in the bacterioferritin was confirmed by EXAFS prior to data collection. Solution of the structure revealed the presence of two heavy metal atom sites in close proximity to each other per monomer. Analysis of the elemental content of the bacterioferritin by µPIXE (Fig 5.7) shows iron, zinc and copper to be present in the sample. The buffer and backing alone were completely clean of iron, zinc and copper. Due to the presence of chloride in the buffer which gave a large X-ray peak next to the sulfur peak, it was not possible to quantify the amounts of iron, zinc and copper present.
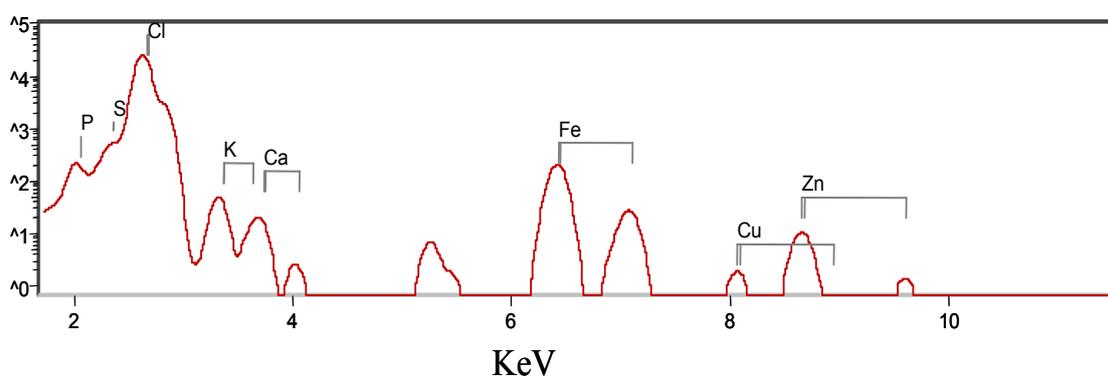
Fig 5.10. Figure showing the µPIXE point spectra for *E. coli* bacterioferritin. Statistical analysis shows iron, copper and zinc to be the only metals present.

Analysis of the X-ray data including the co-ordination geometry of the ions, their B-factors and the absence of strong difference peaks at the K-absorption edge of iron calculated at 10-1.9Å resolution with phases from the refined model, all suggest that zinc is bound to the ferroxidase site. The peak height for the heme iron was 62σ while the peak height of ions at the dinuclear site was in the range of 7-8σ which is comparable with sulfur peaks of 6-9σ. The absence of strong difference peaks suggests that iron is not bound, leaving zinc and copper as potential candidates. Both copper and zinc are capable of binding to the ferroxidase site, zinc however has a much higher affinity to that of copper (K_d, Zn ≈ $1x10^{-7}$ M, Cu ≈ $2x10^{-5}$ M) (Baaghil *et*. *al*. 2002) and given the higher concentration of zinc, the metal sites are likely occupied by zinc. In addition to binding to the ferroxidase site, copper is also thought to bind to low affinity sites on the interior of bacterioferritin (Baaghil *et*. *al*. 2002). Given the levels of copper observed and the inability of copper to displace zinc from

the ferroxidase site it is suspected that copper is present at such low affinity sites, at low occupancy, rather than a mixed copper/zinc species in the ferroxidase site.

5.3.2.3 Structure Validation

The final bacterioferritin structure was analysed with the programs PROCHECK and SFCHECK, to validate bond lengths, bond angles, planar groups, chirality and overall stereochemical quality of the model. The final R-factor for the bacterioferritin model was 18.1% (R_{free} = 22%), in the resolution range 23.8 - 1.91Å. The Cruickshank DPI (Diffraction-component Precision Index) was 0.120Å for the coordinate error of an atom with average B-factor. The stereochemical quality of the model proved to be good with a G-factor of 0.2, which is better than expected (-0.3) for the resolution observed. All main chain and side chain parameters were found to be within or better than the expected range for the resolution (Appendix 7).

A Ramachandran plot (Fig 5.8) was produced using the program PROCHECK to assess the quality of the model with respect to the phi-psi bond angles of the main chain backbone. The plot showed that out of 1728 (non-pro/gly) amino acids from all twelve protein chains, 96.2% of these were present in the most favoured regions, the remaining 3.8% were in the additionally allowed regions, and none were found in the generously allowed or disallowed regions.
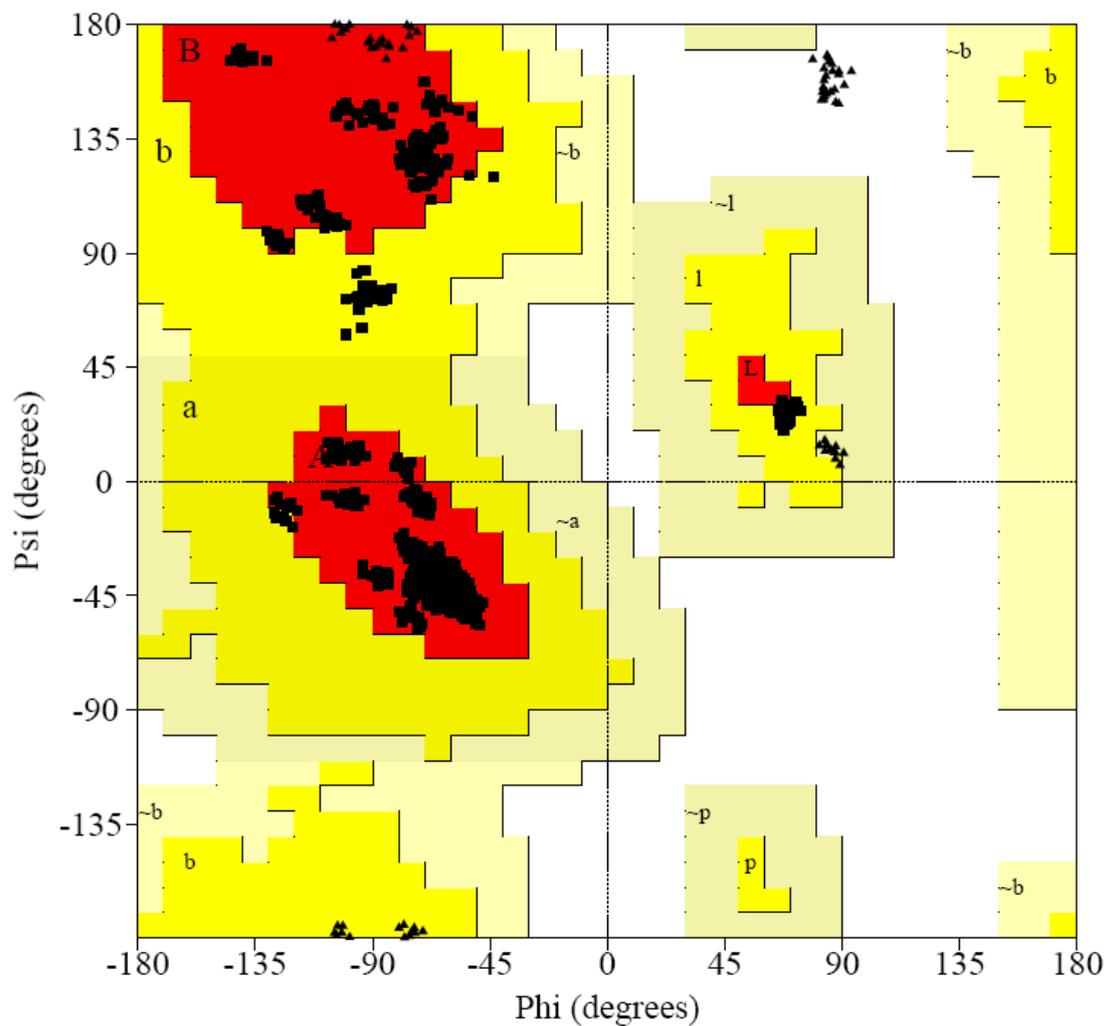
Plot statistics

Residues in most favoured regions [A,B,L]	1662	96.2%
Residues in additional allowed regions [a,b,l,p]	66	3.8%
Residues in generously allowed regions [~a,~b,~l,~p]	0	0.0%
Residues in disallowed regions	0	0.0%
	----	------
Number of non-glycine and non-proline residues	1728	100.0%
Number of end-residues (excl. Gly and Pro)	60	
Number of glycine residues (shown as triangles)	120	
Number of proline residues	12	

Total number of residues	1920	

Fig 5.11.   Ramachandran plot for *E. coli* bacterioferritin showing 96.2% of all residues in most favoured regions of the plot. Figure produced with PROCHECK.

6 TmAKR Structure

6.1 Overall Structure

6.1.1 Monomer

The TmAKR protein consists of 333 amino acid residues (287 and 292 found in the final structure for subunits A and B respectively) which fold into a single domain structure with an α_8/β_8 fold (Fig 6.1), common to all members of the AKR family. In most AKRs the eight β-sheets are arranged into a central barrel surrounded by eight α-helices, in TmAKR the sheets do not however form a fully closed barrel structure. The barrel also does not close in the related (34% identity) AKR11A1 structure (Ehrensberger and Wilson, 2004). This enzyme was shown to have reduced catalytic activity partially due to binding an alternate conformation of NADPH. Reduced activity was not observed in TmAKR (See chapter 3). Each monomer of TmAKR consists of 16.9% β-sheet, 39.7% α-helix and 2.2% 3,10-helix, which form 12 sheets and 10 helices. Eight sheets and eight helices form the α_8/β_8 motif (α1-α8, β1-β8); the four remaining β-strands (S1-S4) form two β-hairpin structures, one at the N-terminal end underneath the barrel, and the other as part of loop 3. The two remaining helices (H1 and H2) are located outside the barrel; H1 is located within loop 7 and H2 in the C-terminal section. Loops are numbered according to the sheet and helix they connect, e.g. loop 1 connects β1 to α1. A topology diagram showing secondary structure numbering is shown in figure 6.2, and a secondary structure/sequence summary is shown in figure 6.3.
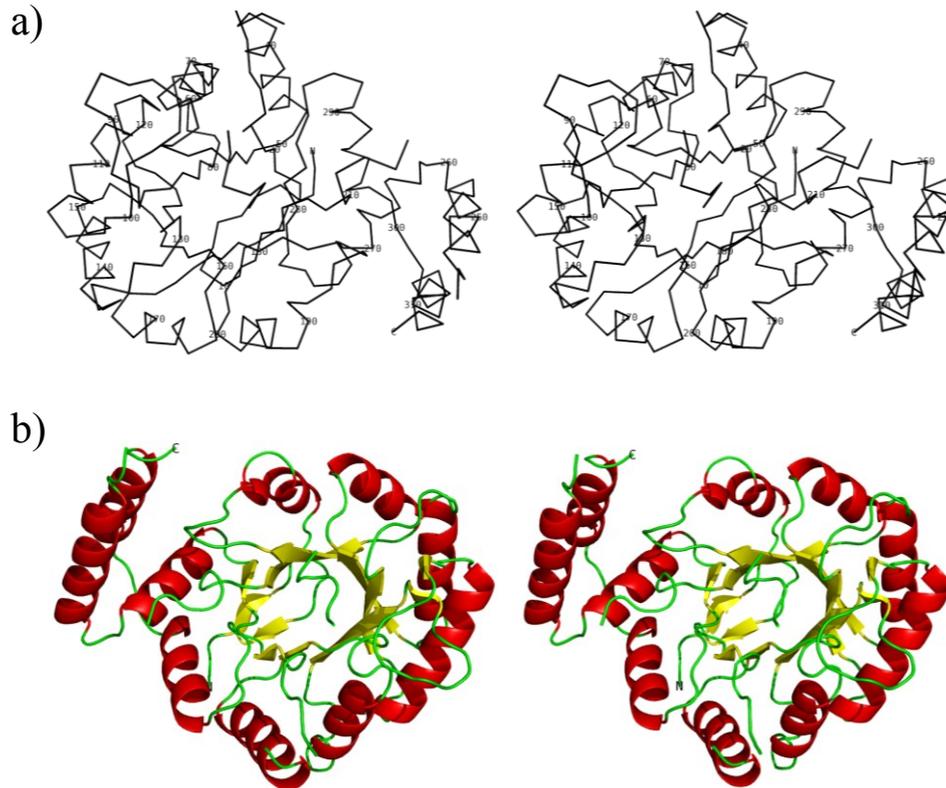
Fig 6.1. Figure showing the overall fold of TmAKR. a) a stereoview of the Cα trace for a single monomer with every tenth amino acid labelled. b) a stereoview of the secondary structure, with α-helices in red, β-sheets in yellow and loop regions in green. (Figure produced with PyMol, Delano Scientific, USA)
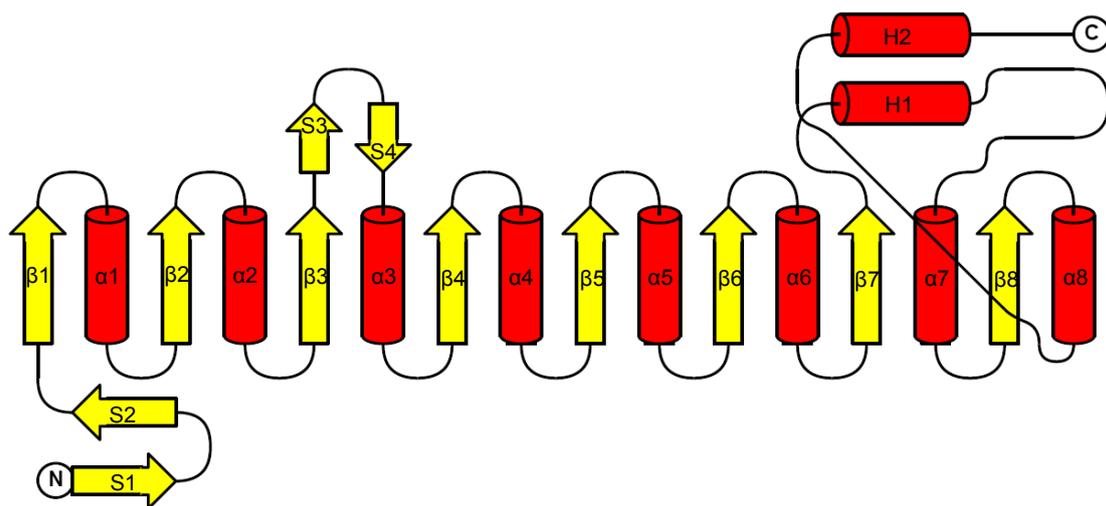
Fig 6.2. A topology diagram of TmAKR. The β-sheets and α-helices from the barrel structure are labelled β1-β8 and α1-α8, respectively. Other β-sheets and α-helices are labelled S1-S4 and H1-H2 respectively. The helices are coloured in red and the sheets in yellow. (Figure created with Topdraw; Bond, 2003)

```
....|....| ....|....| ....|....| ....|....| ....|....|
         10        20        30        40        50

MGIPKRKLGE RGPEVSAIGL GCMRMSFGQK KLPDRKEMIK LIRTAVELGI
....EEETTT TT.EEETTTE ETTTTT       ..HHHHHH HHHHHHHH..
    S1         S2      β₁                    α₁


....|....| ....|....| ....|....| ....|....| ....|....|
        60        70        80        90       100

NFFDTAEVYG PYTNEELVGE ALEPFKGEVV IATKFGFELY EDGRPGWKGL
.EEE.TTTTT TTHHHHHHHH HTTTTTTT.E EEEEE.EE.T TTT......EE
  β₂             α₂                  β₃      S₃         S₄


....|....| ....|....| ....|....| ....|....| ....|....|
       110       120       130       140       150

NSNPEHIKKA VEGSLRRLRV EAIDILYQHR VDPNVPIEEV AGAVKELIEE
...HHHHHHH HHHHHHHH.T TTEEEEEETT TTTTTTHHHH HHHHHHHHH.
     α₃              β₄                α₄


....|....| ....|....| ....|....| ....|....| ....|....|
       160       170       180       190       200

GKVKHFGLCE ASAETIRRAH KVCPVDVVQY EYSMWWRKPE EELLPTCEEL
..EEEEEEET T.HHHHHHHH HH..EEEEEE E.TTTTTGGG TTHHHHHHHH
    β₅         α₅           β₆                    α₆


....|....| ....|....| ....|....| ....|....| ....|....|
       210       220       230       240       250

GIGFVAYSPL GKGFLTGAIG ENSKFDEEDS RSRIPRFQKE NLRENLALVE
..EEEEETTT TTT                              .HHHHHH
   β₇                                             α₇


....|....| ....|....| ....|....| ....|....| ....|....|
       260       270       280       290       300

   R  T                                       N
LLKKIAERKG ATPSQIALAW LLAQKPWIVP IPGTTKLSHL LETIGGAFVE
HHHHHHHHH. ..HHHHHHHH HHHHTTTEEE .......HHHH HHHHGGG...
              H₁               β₈             α₈


....|....| ....|....| ....|....| 
       310       320       330

LTPEELQEIN DALSRIEIKG SRYPEDMEKM TYL
..HHHHHHHH HHHH..
    H2
```

Fig 6.3. The secondary structure summary for TmAKR. The residues shown in green were not observed in the structure and the residues shown in red are mutations from the genome sequence. The original residue coded by the genome sequence is shown above the actual sequence. H – α-helix, E – strand, G – 3,10-helix and T – hydrogen bonded turn.

6.1.1.1 Missing Loops

There are three sections of the protein that are not observed in the final structure, which all correspond to disordered loop regions. The missing sections are from loop 1, loop 7 and the C-terminal loop. The central $\alpha_8\beta_8$ motif is well ordered and has low B-factors whereas the disordered loop regions and helices H1 and H2 which lie outside the central core are associated with higher B-factors and higher disorder (Fig 6.4). Loops 1 and 7 form the 'safety belt' in other AKR structures, interacting with each other and locking the cofactor into place. Loops 1 and 7, along with the C-terminal loop and loop 4, help to form the substrate binding pocket in other AKR structures (Jez *et al.*, 1997). When no substrate or cofactor is bound these loops appear to adopt an alternate disordered conformation. A short section of loop 7 has been located in the active site of the protein interacting with the catalytic tyrosine; the implications of this interaction are unclear and are discussed in Section 6.2.2. The N-terminal methionine is also not observed in the final structure and is likely to be cleaved after translation since the following amino acid is glycine (Sherman *et al.*, 1985).
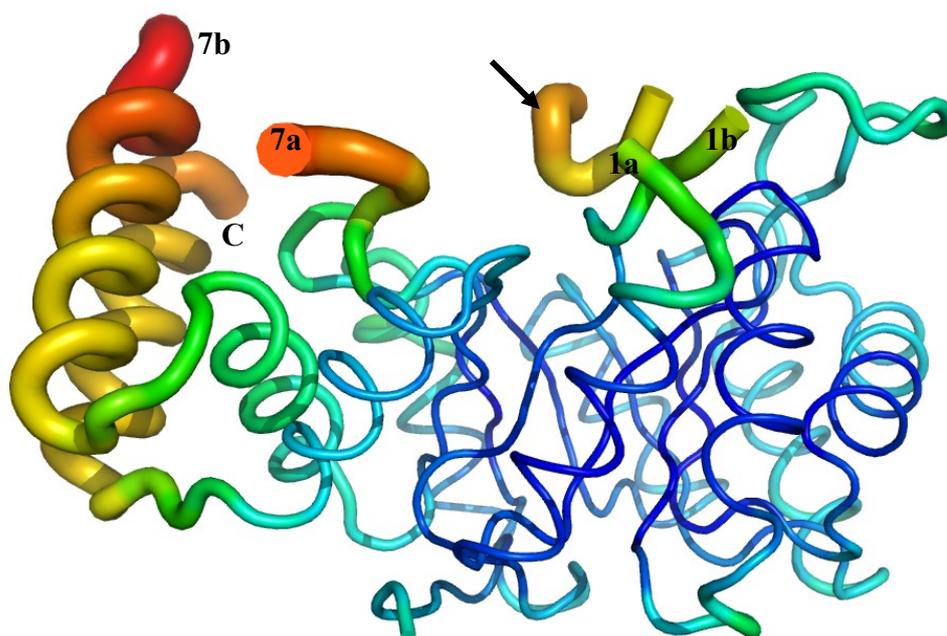
Fig 6.4. Figure showing the Cα-backbone trace of TmAKR, coloured and contoured by the B-factor value. The beginning (a) and end (b) of the missing loops (1, 7 and C) are numbered and a section of loop 7 found in the active site is marked with an arrow. The thin blue lines represent the lower B-factors and more ordered structure, and the heavier orange/red lines represent higher B-factors and a more disordered region. (Figure produced with PyMol, Delano Scientific, USA)

6.1.2 Dimer

6.1.2.1 Crystallographic Dimer

Two subunits were found in the asymmetric unit (Fig 6.5) with one positioned above the other in a head-to-head configuration. The two monomers mainly interact through loop 4 situated near the two-fold non-crystallographic symmetry axis. The symmetry related Pro133 residues from loop 4 interacts on the two-fold axis. The other main interaction is between loop 4 and loop 5 where Asn134 interacts with Thr165 and Arg168, through its amide nitrogen and backbone oxygen respectively. This small numbers of interactions present on the dimer interface mediated through substrate binding loops suggests that this is not the real biological unit of the protein.
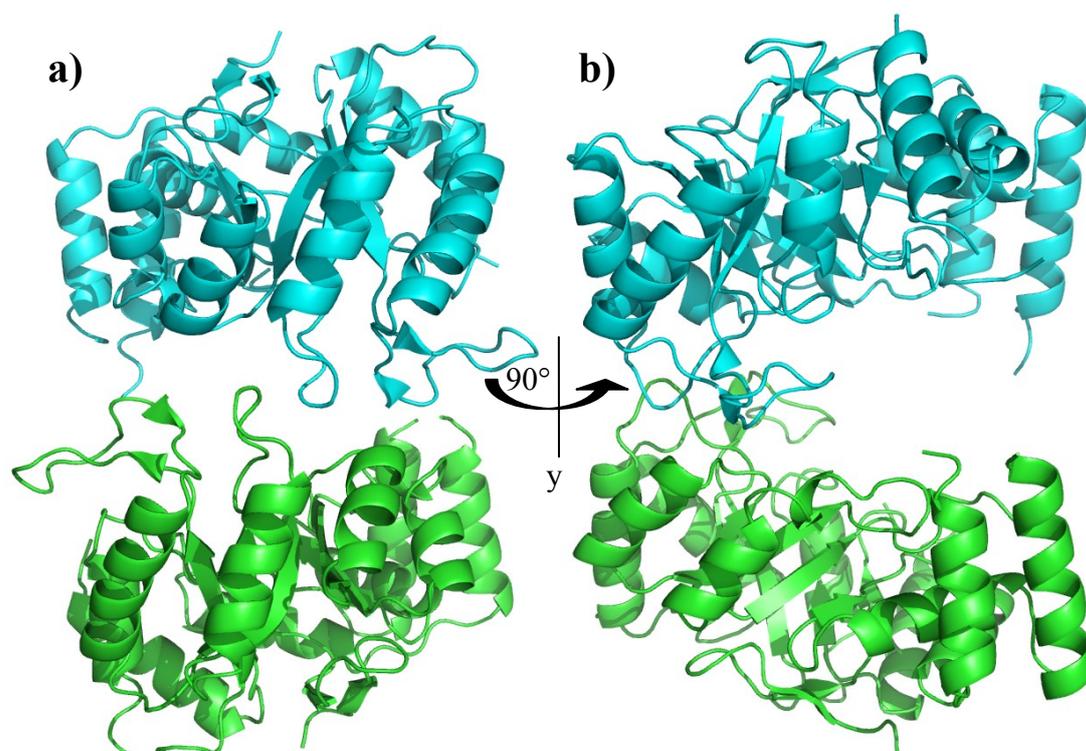
Fig 6.5. The asymmetric unit showing the crystallographic dimer a) viewed down the two-fold non-crystallographic axis, b) as in a), with a 90° rotation around the y-axis. (Figure produced with PyMol, Delano Scientific, USA)

6.1.2.2 Biological Unit

The true biological dimer (Fig 6.6) is centred between the helix two α8s from symmetry related subunits with the interface mainly consisting of hydrophobic interactions. Approximately 2000Å^2 (7.5%) of the solvent accessible surface area is buried upon dimerisation. Helices α1 and α8 form the majority of this contact area which contains a large hydrophobic patch and a number of ionic interactions. Helix α8 forms ionic interactions to loop 7 and helix α1 through hydrogen bonds between Ser288 (helix α8) backbone nitrogen and Gly260 (loop 7) backbone oxygen, and between Phe298 (helix α8) backbone oxygen and Lys40 (helix α1) ε-amino. Helix α1 also forms ionic interactions with the N-terminal glycine (Gly2) amino group through the backbone oxygen and the terminal carboxyl group of Glu47. A salt bridge is also observed between the two subunits of the biological dimer from Asp34 (loop 1) to Arg258 (Helix H1). This salt bridge forms part of a four-residue ion-pair network, linking Asp34-Arg258-Glu305-Lys259 (Fig 6.7) which probably serves to increase the stability of the subunit interaction.

a)

b)

x ⟍ 90°
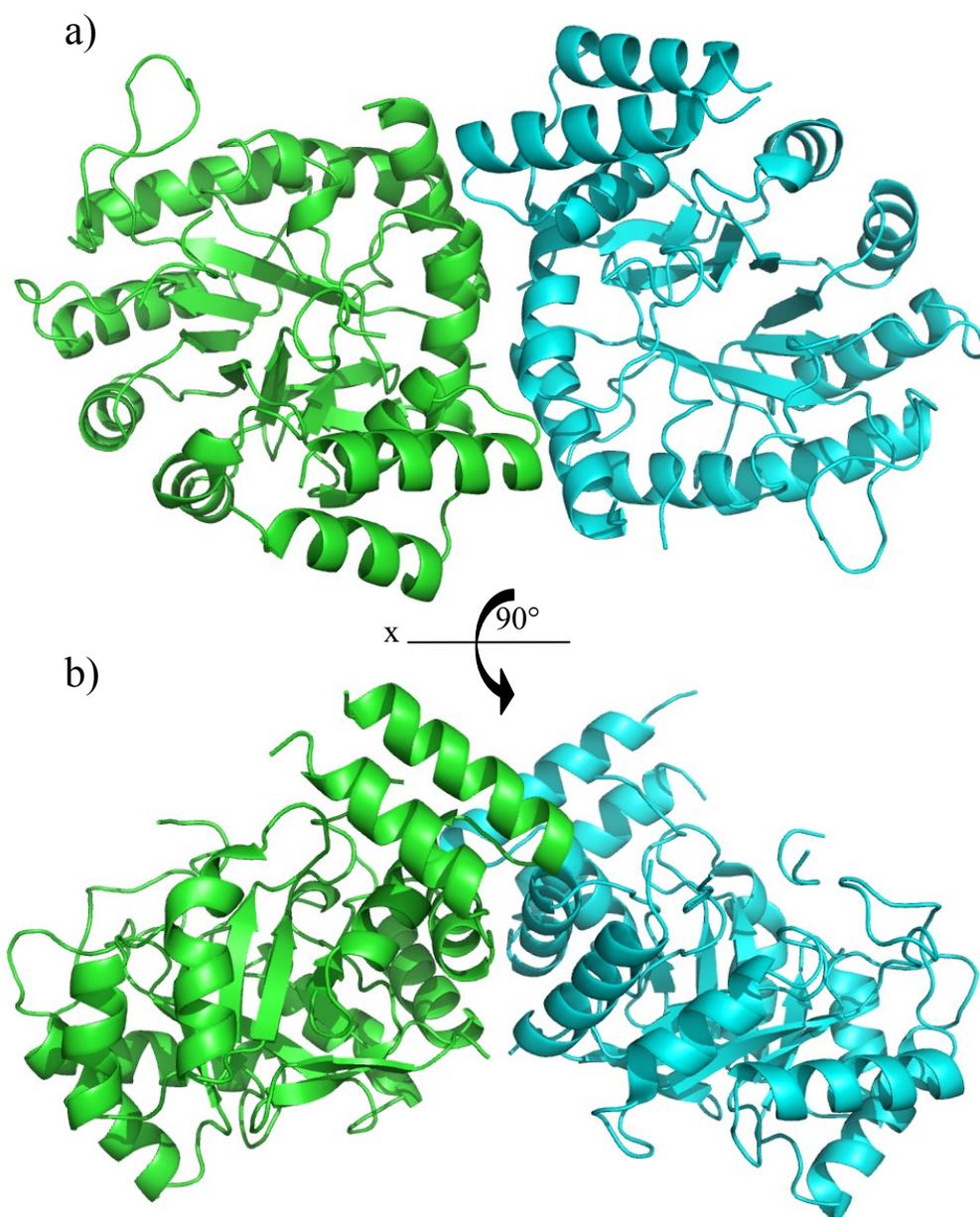
Fig 6.6. Figure showing the biological unit of TmAKR. a) viewed down the two-fold non-crystallographic symmetry axis. b) as in a) with a 90° rotation around the x-axis. (Figure produced with PyMol, Delano Scientific, USA)
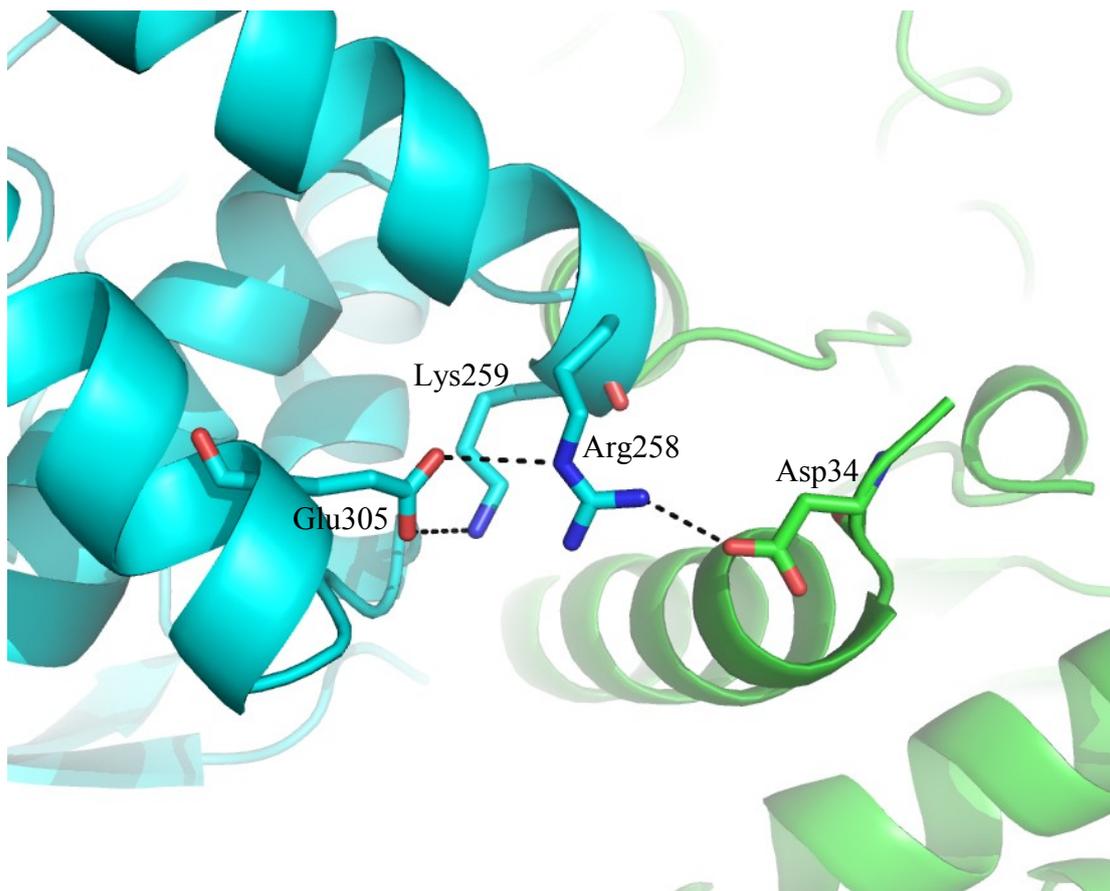
Fig 6.7. The ion-pair network present at the subunit interface in TmAKR between Asp34 and Arg258 on the dimer interface and extending to include Glu305 and Lys259. (Figure produced with PyMol, Delano Scientific, USA)

6.2 Active site

6.2.1 Catalytic tetrad

The catalytic tetrad of tyrosine, lysine, aspartic acid and histidine found in other AKR structures is present within the structure presented here. This tetrad (Fig 6.8) consists of Tyr59, Lys84, Asp54 and His129. The catalytic tyrosine is hydrogen bonded to the lysine, which in turn forms a salt bridge with the aspartic acid. This arrangement serves to lower the pKa of the catalytic tyrosine and enhance the catalytic activity as shown by mutagenesis studies (Tarle *et al.*, 1993; Bohren *et al.*, 1994). The histidine is positioned near to the tyrosine but does not directly hydrogen bond to it. This residue is thought to form an oxyanion hole which may stabilise an oxyanion intermediate. This histidine is often found to hydrogen bond to the catalytic tyrosine in other AKRs via a water molecule. In this structure Glu227 is found positioned in the active site providing the linking hydrogen bonds between His129 and Tyr59. Asp54 is found to form a further hydrogen bond with Gln179 which also hydrogen bonds to Tyr127, extending the proton relay system present in the basic catalytic tetrad to include seven residues. The catalytic significance of Gln179 and Tyr127 are not known and no appropriate mutagenesis studies have been performed. Gln179 is well conserved throughout the AKRs and is often found to interact with the nicotinamide substituent of the NADPH cofactor. Tyr127 is conserved in AKR11A1 (Tyr123) but not in other AKRs which generally have a leucine at this position. Since Tyr127 is four hydrogen bonds from the catalytic tyrosine and is not well conserved, this probably has little effect on catalysis. Gln179 however may affect catalysis but this is more likely to be due to altered cofactor binding rather than catalytic efficiency.
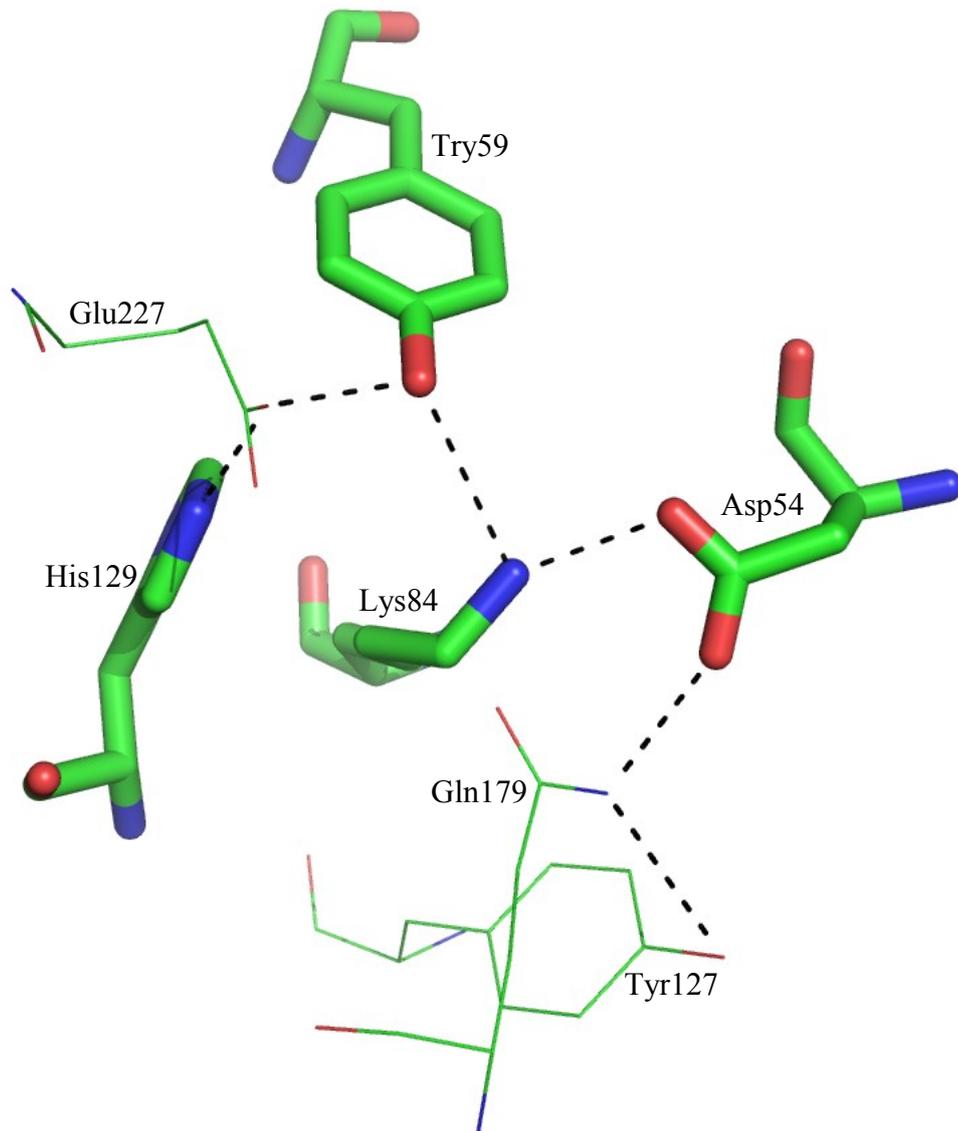
Fig 6.8. The arrangement of residues forming the catalytic tetrad. The tetrad residues are shown as sticks with the other interacting residues shown as lines. The ionic interactions are shown as black dashed lines. Tyr59 is the catalytic residue and Glu227 occupies the site of substrate binding. (Figure produced with PyMol, Delano Scientific, USA)

6.2.2 Active Site Density

A portion of unexplained electron density was found to be present within the active site of TmAKR (Fig 6.9). This unexplained density could not be accounted for by NADPH cofactor or any known substrate of the enzyme. Due to the shape of the density a short section of a polyalanine chain was build into the density and after further refinement putative side chain identities assigned. By comparison of the assigned amino acid sequence with the sequence of the missing loops the unexplained density was identified to be part of the missing loop 7 (224-229). From this loop Glu227 interacts with the catalytic tetrad residues Tyr59 and His129. The carboxyl group of the glutamate occupies a position similar to that which the substrate carbonyl would occupy between the catalytic tyrosine and histidine. The conformation adopted by the glutamate carboxyl would be excluded by NADPH binding due to steric hindrance. Therefore upon cofactor binding the loop 7 would be expected to leave the active site and allow assembly of the substrate binding site to occur. The effect of binding this glutamate at the active site may reduce the conformational entropy of loop 7, and stabilise this loop and the whole protein when no cofactor is bound.
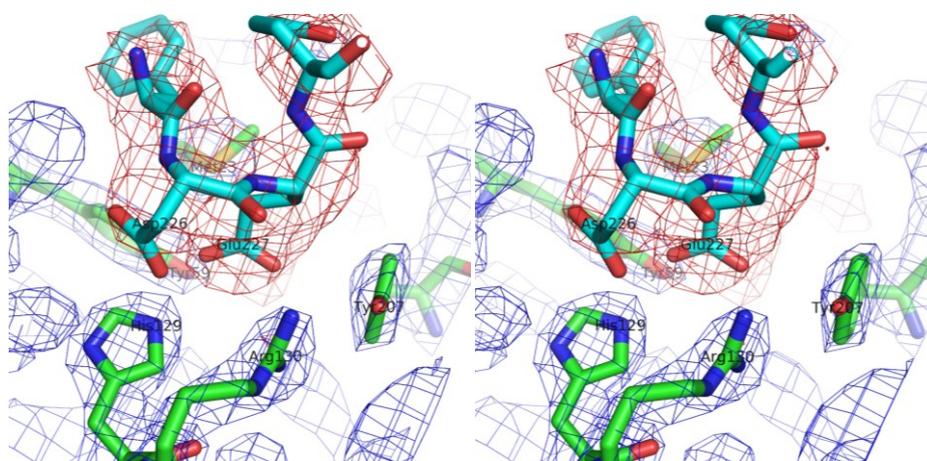
Fig 6.9. A stereoview of the density found within the active site showing the $2F_o$-F_c map in red contoured at 2σ and F_{obs} map in blue contoured at 1.5σ and generated with residues 224-229 omitted from the calculations. The omitted residues are shown in cyan and the neighbouring residues are shown in green. (Figure produced with PyMol, Delano Scientific, USA)

6.2.3 Cofactor binding site

In all AKR structures available the NADPH cofactor binds in a well defined cleft, with the nicotinamide positioned in the active site at the centre of the barrel and the

pyrophosphate positioned across the lip of the barrel between loops 7 and 8. In the TmAKR structure no cofactor was found, and it appears to have been displaced by two sulfate molecules. One of these sulfates is found in a position which would usually be occupied by the pyrophosphate and the other in a position usually occupied by the adenosine 2' monophosphate. These two sulfates are chelated by similar residues to those found in other AKRs interacting with the phosphates of the cofactor. The sulfate in the pyrophosphate position interacts with backbone amides from Lys212 and Leu210 in a similar fashion to most other AKRs. The other sulfate also interacts with two backbone nitrogens (from Lys286 and Thr285), and also with His289. This histidine is usually a well conserved arginine (Jez *et al.*, 1997) but adopts a conformation which should still be able to bind the adenosine 2' monophosphate. Both sulfates form hydrogen bonds with the side chain of Arg24, which adopts a position close to that of Arg270 in the rat 3α-hydroxysteroid dehydrogenase (3α-HSD; AKR1C9) structure, which interacts only with the pyrophosphate bridge and not the monophosphate. Tyr207 is found in a position where it would be able to stack with the nicotinamide ring and orientate it correctly within the active site. The most highly conserved residues in the cofactor binding site are found in similar positions in TmAKR. Asp54 and Gln179 are amongst the most strictly conserved residues for cofactor binding and are also conserved in TmAKR. Asp54 hydrogen bonds to the ribose 2' hydroxyl and Gln179 interacts with the carboxamide of the nicotinamide group. Some other residues which are usually highly conserved across the AKR cofactor binding sites are however not conserved here. Glu160 is strictly conserved as asparagine in all other AKRs and donates a hydrogen bond to the nicotinamide's carboxamide group. Here it is seen in a position rotated away from the cofactor binding site. Arg130 is found instead in a position able to provide a hydrogen bond to the nicotinic carboxamide. Gly283 is usually found as either a lysine or arginine which hydrogen bonds to the pyrophosphate bridge. Arg24 occupies a similar position to this residue and is found hydrogen bonding to both the sulfate molecules. The well defined and highly conserved cofactor binding site across the AKR family allows for the prediction of the potential cofactor contacts within TmAKR (Fig 6.10).

OH

Tyr$_{207}$

Arg$_{130}$

H$_2$N   NH$_2$

O

NH$_2$ |||||||||| O══Gln$_{179}$

NH$_2$

OH |||||||| O

HO ─ Asp$_{52}$

|||||||| HN─Met$_{23}$

OH |||||||| HN─Cys$_{22}$

H$_2$N   NH$_2$ ||||||||||||||| O⁻

HN

Arg$_{208}$─NH |||||||| O⁻

P══O ||||||||| H$_2$O ||||||||| HN─Arg$_{24}$

P══O ||||||||| HN─Gly$_{283}$

HO

O

NH$_2$

Arg$_{286}$

NH$_2$

NH$_2$ ||||||||| O⁻

P

O   O⁻ ||||||||||| H

N

His$_{289}$

OH
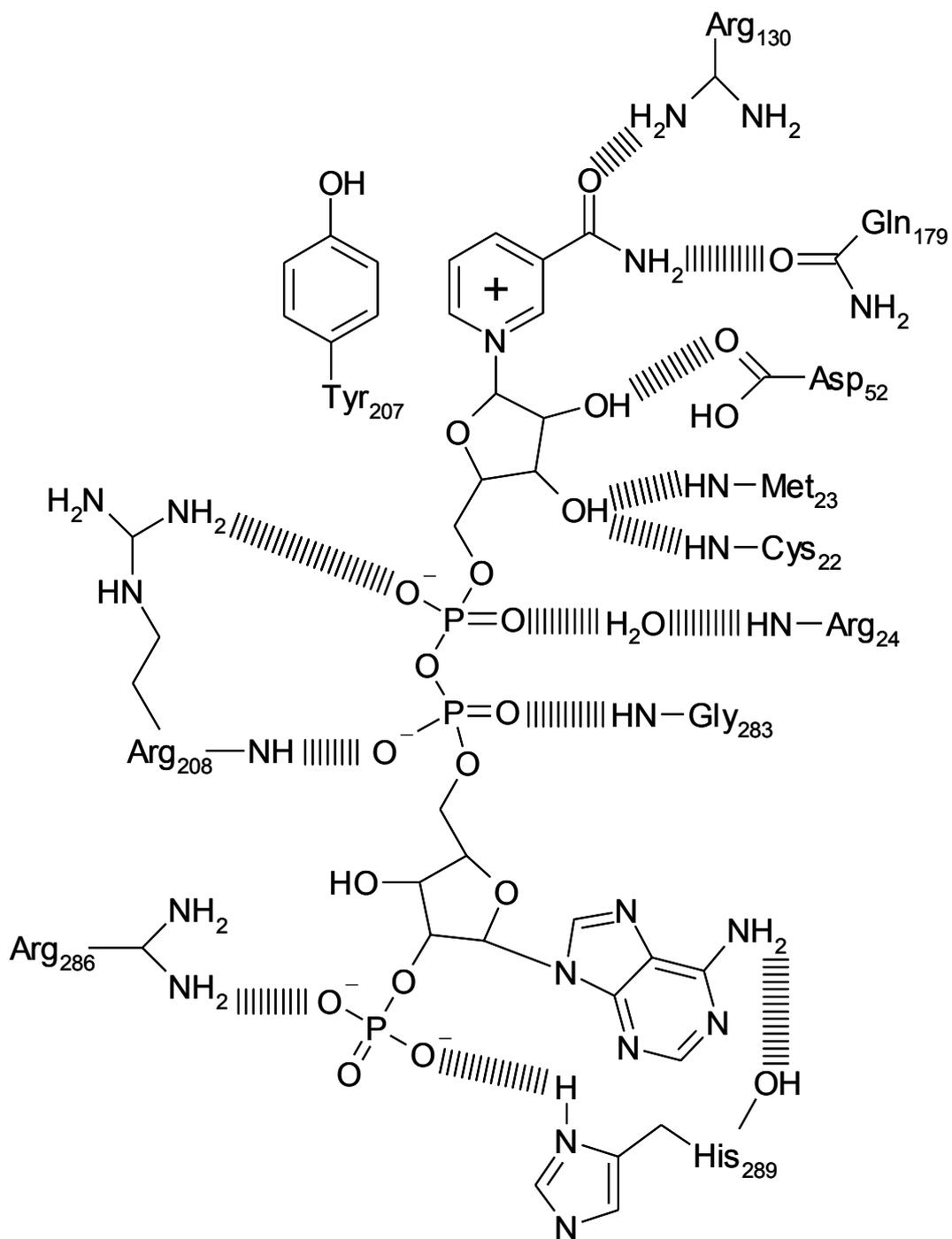
Fig 6.10. A schematic representation of the predicted hydrogen bonds and other significant contacts between NADP$^+$ and TmAKR. Hydrogen bonds shown as dashed bars.

6.3 Thermostability

The enzymes obtained from thermophilic and hyperthermophilic organisms are active and stable at higher temperatures that their mesophilic counterparts. As the number of high resolution structures available increases it is possible to compare similar enzymes from organisms with differing growth temperatures to provide an insight into thermostabilisation mechanisms. No clear set of rules have been found to define how thermophilic enzymes gain their stability at increased temperature. Thermostability is achieved by employing a variety of mechanisms which vary between organisms and enzymes. A range of common mechanisms have been reported to be used and these are discussed in section 1.4.

Most AKRs studied are either from humans and other animals, due to their therapeutic applications, or from yeasts for their ability to convert xylose to xylitol. A number of AKRs from bacterial sources have also been investigated and their structures solved. However the number of bacterial AKR structures (18 structures for 12 unique proteins) is limited compared to other sources. Thermophilic and hyperthermophilic AKR structures are also not well investigated with only a single structure for a hyperthermophilic AKR present in the PDB. This single available structure, also from *Thermot. maritima* (25% identity to TmAKR; 27% identity to AKR11B1) was the product of a structural genomics project (Canaves *et al*., 2004), and has therefore had very little functional characterisation. To allow for identification of thermostabilisation mechanisms utilised by TmAKR, the structure has been compared to AKRs from mesophilic organisms. The other AKR from *Thermot. maritima* (TM1009; PDB: 1VP5) has been included to assess if these two enzymes have evolved to use similar thermostabilisation mechanisms. Three AKR11 family enzymes (AKR11A1, B1 and C1), which represent the structures with the highest identity to TmAKR, have been used as mesophilic examples with which to provide a comparison. The AKR11A1 enzyme from *B. subtilis* has the highest similarity to TmAKR (34% identity) but has reduced catalytic activity and binds NADPH in an alternate conformation; therefore comparision with AKR11B1 also from *B. subtilis* (32% identity) has been concentrated upon. AKR11C1 is from *B. halodurans* (31% identity to TmAKR).

6.3.1 Amino Acid Composition

The overall amino acid composition of a protein was thought to correlate with thermostability. Genome sequencing projects have shown however that there are no global trends, but thermophilic proteins may show a preference for certain amino acids. In general these are only minor trends and are organism dependant. The amino acid composition of the two *Thermot. maritima* AKRs (TmAKR, TM1009) and the three AKR11 family members studied are shown in table 6.1.

	TmAKR	TM1009	AKR11A1	AKR11B1	AKR11C1
Growth Temperature	80°C	80°C	25-35°C	25-35°C	25-35°C
Ala (A)	6.60%	5.90%	6.70%	7.80%	7.10%
Arg (R)	6.30%	5.90%	4.50%	5.10%	7.40%
Asn (N)	3.00%	3.50%	6.70%	3.60%	2.70%
Asp (D)	3.00%	4.20%	7.10%	6.30%	6.10%
Cys (C)	1.20%	1.00%	0.00%	0.30%	0.70%
Gln (Q)	2.10%	4.50%	2.20%	3.90%	5.40%
Glu (E)	12.90%	10.50%	9.00%	9.00%	9.10%
Gly (G)	8.70%	5.90%	6.70%	7.20%	7.10%
His (H)	1.50%	2.10%	2.60%	1.50%	2.40%
Ile (I)	6.90%	8.40%	7.40%	8.10%	6.40%
Leu (L)	9.90%	6.60%	9.90%	9.00%	11.10%
Lys (K)	7.50%	7.30%	8.70%	6.00%	5.70%
Met (M)	2.10%	3.50%	1.00%	2.40%	1.70%
Phe (F)	3.60%	4.50%	5.40%	2.70%	1.70%
Pro (P)	6.00%	4.20%	4.50%	5.10%	3.70%
Ser (S)	4.20%	3.50%	3.80%	4.20%	5.10%
Thr (T)	3.90%	3.80%	4.50%	7.80%	4.00%
Trp (W)	1.50%	1.70%	0.30%	2.10%	1.30%
Tyr (Y)	2.70%	4.20%	3.20%	3.60%	4.70%
Val (V)	6.30%	8.40%	5.80%	4.20%	6.70%
Charged	31.20%	30.00%	31.90%	27.90%	30.70%
Uncharged polar	5.10%	8.00%	8.90%	7.50%	8.10%
Aromatic	7.80%	10.40%	8.90%	8.40%	7.70%

Table 6.1. The amino acid composition of the two hyperthermophilic *Thermot. maritima* AKRs (TmAKR and TM1009) and the three mesophilic AKR11 family enzymes.

The TmAKR enzyme has a reduced proportion of uncharged polar amino acids than the mesophilic enzymes which may reduce the occurrence of deamidation. This reduction in uncharged polar residues is often observed in hyperthermophilic proteins and is often accompanied by an increase in charged amino acids which may increase the opportunities to form ionic bonds; in TmAKR only an increase in glutamate is observed. An increase in proline and glycine residues is also observed. Increasing the proportion of proline may provide an increase in stability by increasing rigidity of the protein particularly in loop regions. Whilst increasing the glycine content may serve to release conformational strain in certain parts of the structure. The TM1009 enzyme has a differing amino acid composition to TmAKR. TM1009 has increased amounts of aromatic and hydrophobic residues (particularly valine) compared to both TmAKR and the AKR11s. Other trends in the TM1009 enzyme compared with the mesophilic enzymes are for more glutamate and methionine along with reduced aspartate. These alterations in composition may have more of an evolutionary significance rather than an effect on thermostability. Interestingly both the *T. maritima* enzymes have increased cysteine content; thermophilic enzymes usually contain fewer cysteine residues due to their reactivity, and susceptibility to auto-oxidation.

6.3.2 Structural Features

The α-helices within a protein carry an inherent dipole where the amino terminus carries a partial positive charge and the carboxyl terminus carries a partial negative charge. Stabilisation of these charges by placing oppositely charged residues at either end may stabilise the helices and prevent unfolding of the secondary structure. TmAKR has more of its helices stabilised at both N and C termini compared to the three AKR11 structures and TM1009 (Table 6.2), suggesting that this may be a mechanism of stabilisation which the TmAKR enzyme utilises. In contrast to TmAKR the other *Thermot. maritima* AKR, TM1009, has a similar number of termini stabilised to the AKR11s. In addition to having more helices stabilised, TmAKR has more N-termini stabilised by the presence of proline residues. This correlates well with the increased presence of proline in the primary sequence of the enzyme. Other additional prolines occur near to the ends of β-sheets, presumably reducing flexibility and conformational entropy and therefore reducing the likelihood of these secondary structural elements unfolding.

Helices stabilised	TmAKR	AKR11A1	AKR11B1	AKR11C1	TM1009
At N-terminus	9	7	6	7	7
At C-terminus	4	0	2	3	3
Total	13	7	8	10	10
By Proline	5	3	2	2	2

Table 6.2. A table showing the α-helix stabilisation with the number of N-termini, C-termini and total termini of the α-helices that are stabilised by an opposite charged residue, with the occurrence of proline residues at the N-termini also shown.

Most microbial AKRs are found as monomers in solution with the exception of some yeast xylose reductases (Yokoyama *et al*., 1995). Both of the *Thermot. maritima* AKRs are dimeric in nature which may contribute towards their increased stability. The TmAKR was shown to be dimeric by gel filtration chromatography (Section 2.3.2) and TM1009 is predicted to be dimeric by the EBI Protein Quaternary Structure (PQS) server (www.ebi.co.uk; CCP4, 1994; Vriend, 1990). Both enzymes form their dimers with quite different interfaces, TmAKR has its dimer interface centred around helices α1 and α8, with an angle of approximately 90° created between the monomers (Fig 6.6). The TM1009 enzyme has its dimers rotated 180° relative to each other (Fig 6.11) with the dimer interface mainly composed of helices α4 and α5. The two dimer interfaces are also varied in their nature with the TmAKR interface being mainly composed of hydrophobic interactions with all water excluded and only a small number of ionic interactions (see section 6.1.2.2). In contrast the TM1009 dimer interface contains two small hydrophobic patches and many hydrogen bonded interactions which are mostly water mediated (4 direct hydrogen bonds). All three of the AKR11 members studied are monomeric in solution.

a)

b)

x ———— 90°
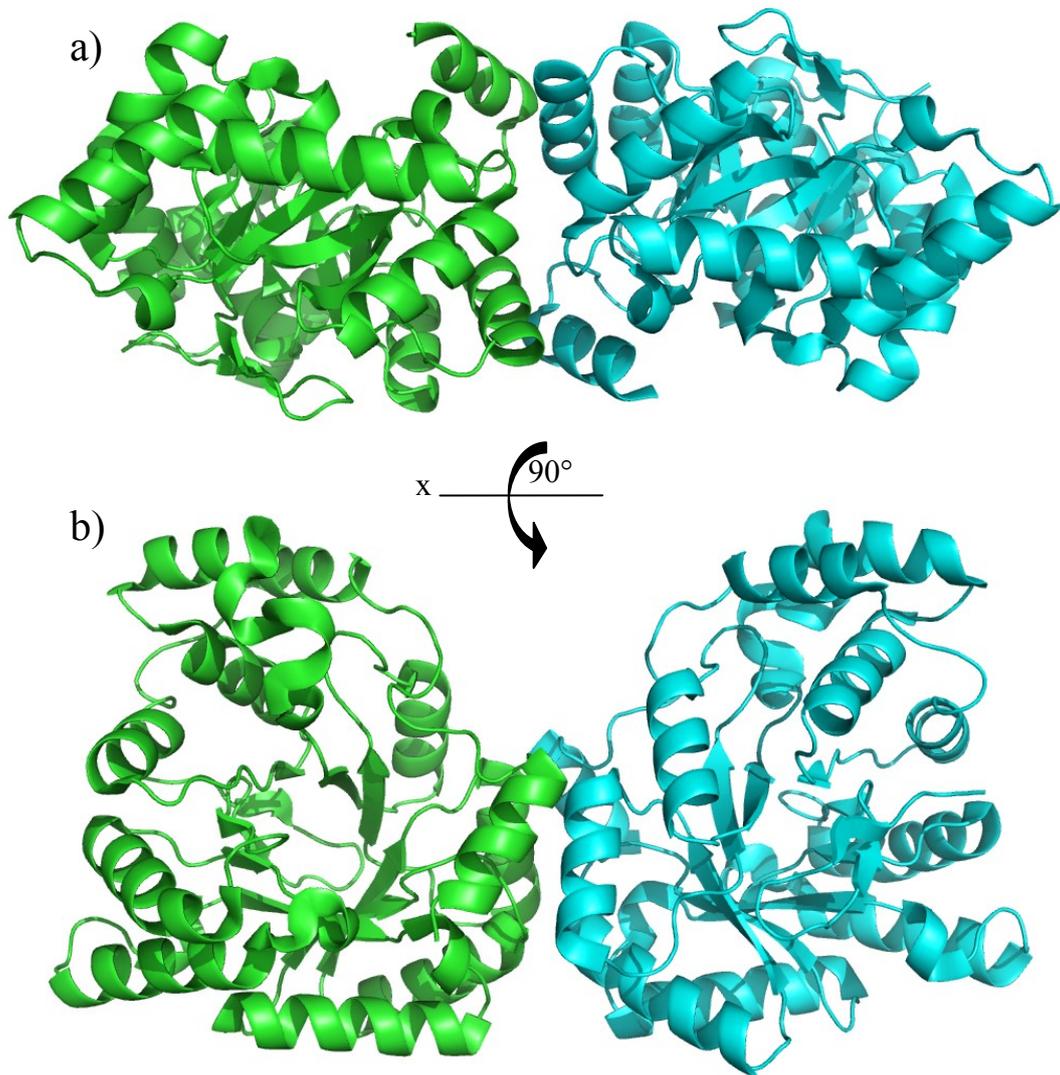
Fig 6.11. Figure showing the TM1009 dimer as predicted by the EBI PQS server (www.ebi.co.uk). a) the view down the non-crystallographic two-fold axis on the dimer interface, b) as in a) with a 90° rotation about the x-axis. (Figure produced with PyMol, Delano Scientific, USA).

Ion-pair networks are often found to have a higher occurrence and to be larger in thermophilic proteins. The TmAKR monomer possesses two more ionic interactions than AKR11B1, and both contain two networks with three residues suggesting that an increase in the size and occurrence of ion-pair networks is not a major contributory factor to the stabilisation of TmAKR. The ion-pairs present in TmAKR are however more often found between adjacent secondary structural elements, thereby connecting

two parts of the structure together which are distant in the primary sequence, which may offer a stabilising effect (Perutz and Raidt, 1975). If we consider the TmAKR dimer then one of the three residue networks is extended to four residues through the dimer interface. This four residue ion-pair network is further extended through hydrogen bond interactions to provide N-terminal helix stabilisation to helix H2 in one subunit and to helix α1 in the other subunit (Fig 6.12).
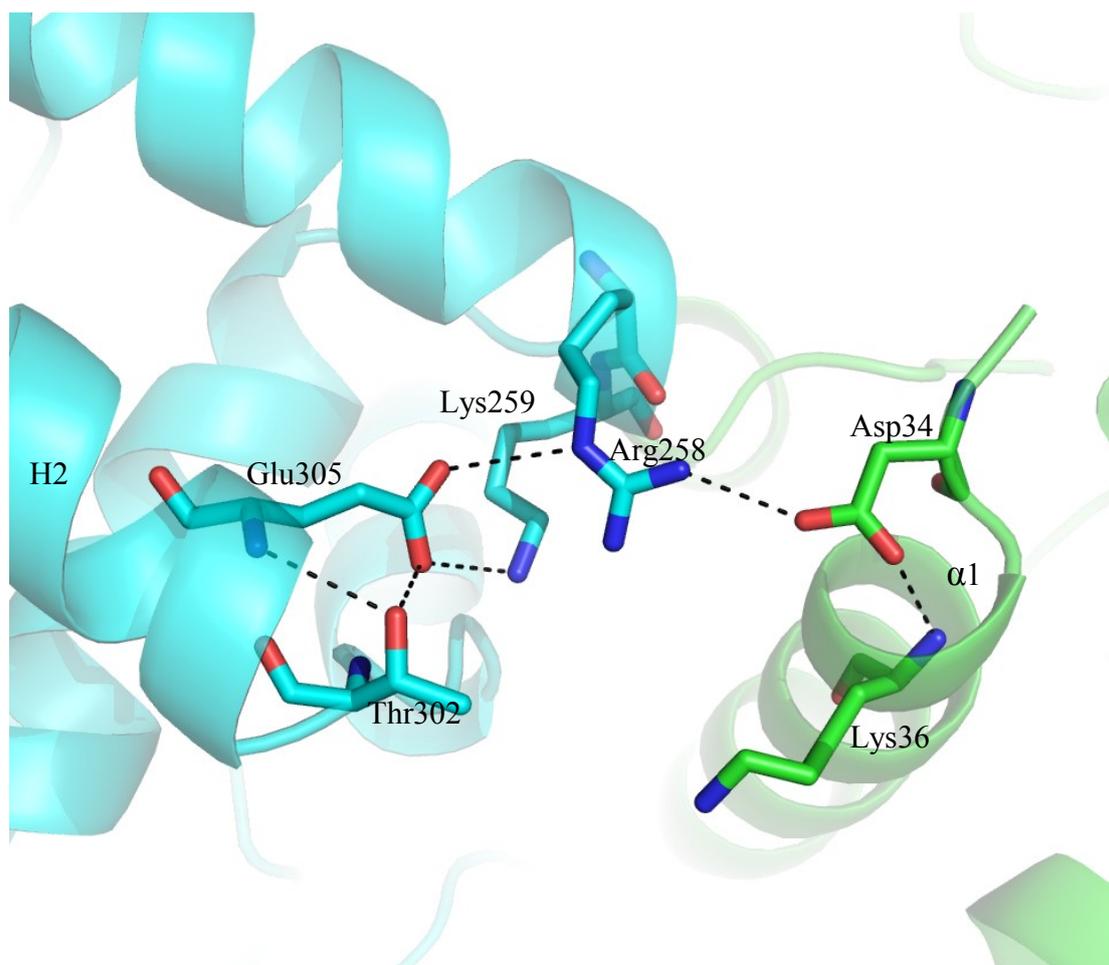
Fig 6.12. Figure showing the ion-pair network at the dimer interface in TmAKR consisting of four residues (Asp34-Arg258-Glu305-Lys259), which interacts with the N-termini of helix α1 and helix H2 to provide stabilisation to the helix dipoles. (Figure produced with PyMol, Delano Scientific, USA).

In TM1009 there is the same overall number of ion-pair interactions compared with TmAKR but more of these are found in larger networks. Two networks of three residues are found in addition to a four residue network and a five residue network. Unlike TmAKR none of these ion-pair networks cross the dimer interface. This suggests that TM1009 utilises larger ion-pair networks as a general stabilising mechanism, whereas TmAKR utilises more single ions-pairs separated in sequence to provide a stabilising effect in addition to using an ion-pair network to stabilise the dimer interface.

The two *Thermot. maritima* AKRs appear to gain their thermostability by differing mechanisms. TmAKR utilises increased helix dipole stabilisation, long distance ion-pair interactions and dimer formation to achieve its increased stability. In contrast the TM1009 enzyme predominantly utilises increased size of ion-pair networks in addition to dimer formation to achieve its increased stability. The analysis of TM1009 thermal stabilisation has been carried out relative to TmAKR and the AKR11A1, B1 and C1 enzymes, but TM1009 is most closely related to the AKR5 family and comparison with this family may further reveal which mechanisms are utilised to achieve stabilisation. However the analysis performed here does suggest that the *Thermot. maritima* AKRs have evolved distinctly different thermostabilisation mechanisms.

7 Bacterioferritin Structure

7.1 Introduction

Bacterioferritin from *E. coli* is an iron storage protein consisting of 24 identical subunits (18 kDa each) with 432 symmetry, which form a roughly spherical structure approximately 120Å in diameter, with an 80Å hollow core. Iron is deposited within the core by the ferroxidase reaction which occurs in 3 distinct kinetic steps as determined by EPR spectroscopy (Le Brun, 1993). 1) the binding of two Fe(II) ions per subunit; 2) the fast oxidation of Fe(II) to Fe(III) by reduction of O_2 or H_2O_2; 3) slow core formation phase where iron is deposited as an insoluble ferric oxyhydride phosphate mineral. The presence of phosphate during core formation has a significant enhancing effect upon phase 3 by increasing the transfer rate of electrons from the core surface to the ferroxidase centre (Aitken-Rogers *et al*., 2004).

As well as the heme binding bacterioferritin, *E. coli* also possesses another iron storage protein in the form of the non-heme binding ferritin. Ferritin is an iron storage protein with a similar overall architecture to bacterioferritin but lacking heme. It is made up of 24 identical subunits of approximately 18 kDa, binds two non-heme metals per subunit and also exhibits 432 point group symmetry. It has been suggested that the ferritin and bacterioferritin monomers are the bacterial equivalents of the H and L chains of the mammalian ferritin system (Stillman, 2001).

7.1.1 Ferroxidation Mechanism

The mechanism of core formation and the contribution which the ferroxidase centre makes is not clear. A variety of mechanisms have been suggested; one where iron is oxidised at the ferroxidase centre followed by release to the core where nucleation and growth can occur. Evidence for this mechanism comes from the structures of *Azotobacter vinlandii* (Liu *et al*., 2004; Swartz *et al*., 2006) and *Desulfovibrio desulfuricans* (Macedo *et al*., 2003) bacterioferritins where there is a concerted movement of His130 (His59 in *D. desulfuricans*) allowing release of Fe(III) into the core. This is supported by the observation of partial occupancy in the corresponding (FE2) iron site. In the *E. coli* structures described to date, this movement of His130 is not observed, however in the mixed iron/zinc structure (van Eerde *et al*., 2006) and the manganese bound structures (Frowlow & Kalb, 2001) partial occupancy of the

FE2 site (ligated by His130) was observed and was attributed to a similar mechanism as seen in *A. vinlandii* and *D. desulfuricans* bacterioferritins.

An alternative mechanism has been proposed where the iron at the ferroxidase site acts like a cofactor. Fe(II) is oxidised to Fe(III) at the ferroxidase site by reduction of O_2 or H_2O_2. Fe(II) which has already entered the core (via alternate channels in the protein) is oxidised either at nucleation sites on the protein inner surface or on the growing core surface, where electrons are channelled back through the ferric oxyhydrite phosphate core to the ferroxidase site where Fe(III) is reduced to Fe(II) and the cycle can begin once more. Therefore by this mechanism oxidation of iron and reduction of oxygen/peroxide become spatially separated. Evidence for this mechanism comes from the requirement for a functional ferroxidase site throughout the whole core formation process and the dependence of the rate of core formation on the surface area of the growing core, an observation which indicates the involvement of the core surface in the catalytic mechanism (Baaghil *et al*., 2003; Lewin *et al*., 2005).

7.1.2 Heme

The bacterioferritins bind 12 heme molecules per icositetramer but little is known about their function. The heme is ligated by two Met52 residues from neighbouring subunits within what appears to be a symmetrical hydrophobic pocket. The neighbouring Glu51 residue forms a ligand to each of the metals within the ferroxidase centre thereby providing the potential for interaction between the two metal centres. Mutation of the Met52 residues to histidine or leucine prevented heme binding and showed that heme was not required for iron accumulation (Andrews, 1995). These mutants accumulated more iron than their heme containing counterparts suggesting that the heme may regulate iron release from the core. Furthermore when horse spleen ferritin is loaded with heme, iron reduction and release is accelerated by increasing the rate at which electrons are delivered to the core (Kadir *et al*., 1992)

7.1.3 Role in Oxidative Stress

In comparison to the ferritins, the primary role of bacterioferritins may not be for iron storage and they may play a role in oxidative stress resistance. Evidence comes from the time of expression of ferritins and bacterioferritins (Andrews, 1998), where as ferritins are expressed during the growth phase, bacterioferritins are expressed in the stationary phase. This is consistent with ferritin having a role in general house

keeping and growth function, where as the late expression of bacterioferritin points to a role in regulation of stress.

Ferritin knock-outs have shown a reduction in intracellular iron and a decrease in iron-restricted growth, pointing to an iron storage function. Whereas knock-out studies of bacterioferritin failed to show any phenotype relating to iron storage (Andrews *et al*., 2003). This lack of a phenotype suggests an alternate or dual role for bacterioferritin.

The mechanism of bacterioferritin also points to a role in oxidative stress resistance (Bou-Abdallah *et al*., 2002). By utilising pairwise oxidation of Fe(II) in the ferroxidase centre, the bacterioferritin prevents the production of harmful radicals from the Fenton reaction (Eq. 7.1). The ferroxidation mechanism is capable of removing the potentially harmful combination of Fe(II), O_2 and H_2O_2 from the cell. Given O_2 as an oxidant the bacterioferritin will produce H_2O_2 as an intermediate (Eq. 7.2), but this is not released into the bulk solvent and is rapidly utilised as an oxidant by a neighbouring ferroxidase site (Yang *et al*., 2000) (Eq. 7.3).

$$Fe^{2+} + H_2O_2 \longrightarrow Fe^{3+} + HO^- + HO \qquad \text{Eq. 7.1}$$

$$[Fe(II)_2\text{-}P] + O_2 + H_2O \longrightarrow [Fe(III)_2O\text{-}P] + H_2O_2 \qquad \text{Eq. 7.2}$$

$$[Fe(II)_2\text{-}P] + H_2O_2 \longrightarrow [Fe(III)_2O\text{-}P] + H_2O \qquad \text{Eq. 7.3}$$
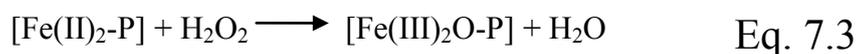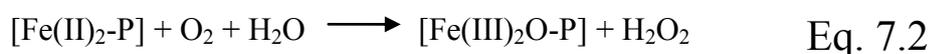
$[Fe(II)_2\text{-}P]$ represents iron bound to the ferroxidase site.

By this mechanism of pair wise oxidation and subsequent elimination of intermediates, the net reaction (Eq. 7.4) gives two oxidised ferroxidase centres with no net release of by-products. In comparison ferritins have been shown to release some of the peroxide intermediate into the bulk solvent, suggesting removal of reactive oxygen species is not as functionally important for the ferritins as it is for bacterioferritins (Lewin *et al*., 2005).

$$2[Fe(II)_2\text{-}P] + O_2 \longrightarrow 2[Fe(III)_2O\text{-}P] \qquad \text{Eq. 7.4}$$
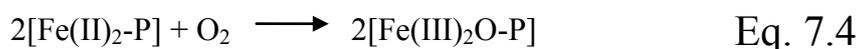
7.1.4 Previous Structures

The X-ray structure of *E. coli* bacterioferritin has previously been solved (Frolow *et al*., 1994; Dautant *et al*., 1998) to 2.9Å (PDB: 1BCF, 1BFR) and more recently to 2.5Å (van Eerde *et al*., 2006; PDB: 2HTN). The arrangement of the subunits exhibits 432 point group symmetry, with channels at the 4-fold and 3-fold axis connecting the inner core to the outside environment (Dautant *et al*., 1998). Previous structures showed the heme binding in a symmetrical hydrophobic pocket located on the 2-fold axes between two symmetry related subunits. Also two metal ions are shown binding to each monomer at the ferroxidase centre and have previously been assigned as manganese (Frolow *et al*., 1994; Dautant *et al*., 1998), or as a mixture of iron and zinc (van Eerde *et al*., 2006).

7.2 X-ray Structure

7.2.1 Overall structure

7.2.1.1 Monomer

Each subunit of the *E. coli* bacterioferritin consists of 158 amino acids which fold up into an all α-helical structure. The monomer (Fig 7.1) consists of a 4 helix bundle with a short 5th helix at the C-terminal end. The 4 helix bundle is composed of 2 anti-parallel helix pairs AB and CD linked by a loop L, which passes over the top of the 4-helix bundle and forms the back of the heme binding site. The helices are arranged so that A and C are paired together on the external face and B and D are paired together on the internal face below A and C. A total of 78% of the structure is composed of α-helices. A summary of the secondary structure is provided in figure 7.2. γ-turns are found between helices A and B (residues 36-38) and helices C and D (residues 112-114). Between helices B and C is a type IV β-turn (residues 70-73), which leads into a γ-turn (residues 72-74). Each monomer binds two metal atoms in the centre of the 4 helix bundle, and contributes one ligand towards chelation of the heme.
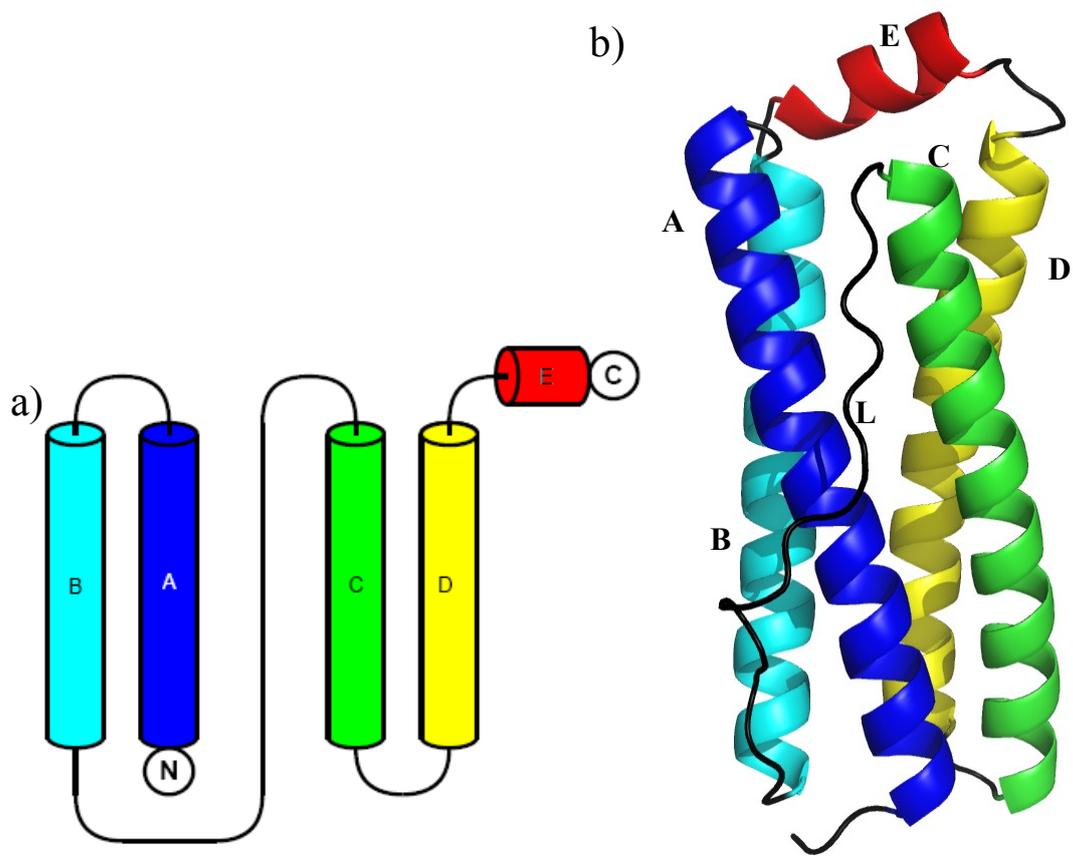
Fig 7.1a) Topology diagram of the *E. coli* bacterioferritin monomer (figure produced using Topdraw, Bond, 2003). b) Cartoon representation of bacterioferritin monomer (figure produced using PyMOL, Delano Scientific USA). Helices (A-E), coloured A – Blue, B – Cyan, C – Green, D – Yellow and E – Red.

```
....|....| ....|....| ....|....| ....|....| ....|....|
        10        20        30        40        50
MKGDTKVINY LNKLLGNELV AINQYFLHAR MFKNWGLKRL NDVEYHESID
....HHHHHH HHHHHHHHHH HHHHHHHHHH HHHH.TTHHH HHHHHHHHHH
          α1


....|....| ....|....| ....|....| ....|....| ....|....|
        60        70        80        90       100
EMKHADRYIE RILFLEGLPN LQDLGKLNIG EDVEEMLRSD LALELDGAKN
HHHHHHHHHH HHHH.....T TTTT...... ..HHHHHHHH HHHHHHHHHH
 α2                                          α3


....|....| ....|....| ....|....| ....|....| ....|....|
       110       120       130       140       150
LREAIGYADS VHDYVSRDMM IEILRDEEGH IDWLETELDL IQKMGLQNYL
HHHHHHHHHH .TTHHHHHHH HHHHHHHHHH HHHHHHHHHH HHHH.HHHHH
                        α4                   α5


....|...

QAQIREEG
H.......
```

Fig 7.2. Structure sequence summary of *E. coli* bacterioferritin. Showing helices (H) and turns (T). Helices labelled α1 – α5. Residues coloured green were not observed in the structure. (Figure produced using Bioedit; Hall, 1999)

7.2.1.2 Dimer

The dimer (Fig 7.3) consists of two monomers in a head to tail arrangement centred on the local two-fold axis. The subunit interface for the dimer is formed by helices A and B from both symmetry related subunits and includes the heme binding site in the centre of the interface. In addition to the heme binding site the interface consists of a number of salt bridges and hydrophobic patches. Ionic interactions are found connecting the two loops from symmetry related subunits, with the carboxamide of Gln72 interacting with Leu77 and Gly75 backbone nitrogen and oxygen respectively. Also holding the two loops together is a hydrophobic interaction between two Leu74s positioned on the two-fold axis. The loop L also interacts with helix A from the opposite monomer through the backbone oxygen of Leu71 hydrogen bonding with the ND2 of Asn23. There are also salt bridges formed between helices A and B from opposite monomers, the terminal oxygens of Asp56 (helix B) form salt bridges with the NE and NH2 of Arg30 (helix A) as does the OE2 of Glu60 (helix B) with the NZ of Lys33 (helix A). These interactions between helices A and B are located together,

where Glu60 also has the potential to form a salt bridge with Arg30, providing a strongly interacting site between the two monomers. There is also a large hydrophobic patch located either side of the heme binding site composed of Phe26 and Leu27 (helix A), Ile59 and Leu63 (helix B) and Pro69, Leu71 and Leu77 (loop L), further serving to bring together helices A and B and loop L and tighten the interaction of the dimer interface. Due to its symmetrical nature all the interactions discussed (except Leu77-Leu77) occur twice on each dimer interface.
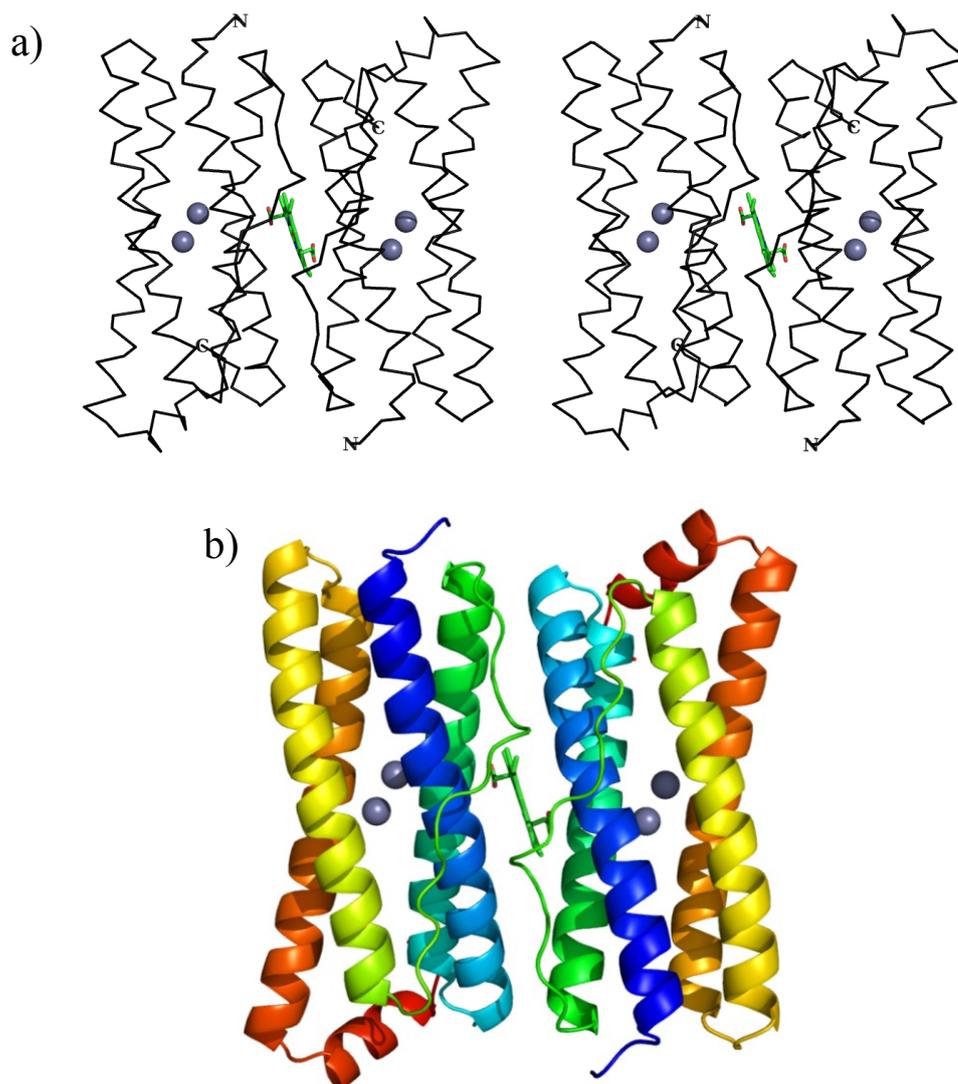
Fig 7.3. Dimer of *E. coli* bacterioferritin, viewed down local two-fold non-crystallographic symmetry axis, shown as: a) stereo view of Cα trace with Met1 and Glu157 labelled and b) cartoon representation coloured from N-terminus (Blue) to C-terminus (Red). Zinc ions are shown as grey spheres and heme as stick representation. (Figure produced using PyMOL, Delano Scientific USA).

7.2.1.3 Icositetramer (24mer)

The biological unit of bacterioferritin consists of 24 subunits (icositetramer), or alternatively can be thought of as 12 dimers. The icositetramer forms a roughly spherical shape with a diameter of approximately 120Å containing an 80Å hollow core, where Fe(III) may be deposited. The bacterioferritin molecule shows extensive non-crystallographic symmetry exhibiting 432 point group symmetry. There are 12 two-fold axes (Fig 7.3b), 3 four-fold axes (Fig 7.4), and 4 three-fold axes (Fig 7.5). The two-fold axes are found on the dimer interfaces where heme binding occurs, the three-fold and four fold axes lie on channels thought to connect the internal and external environments and allow metal atoms to pass through to the core.
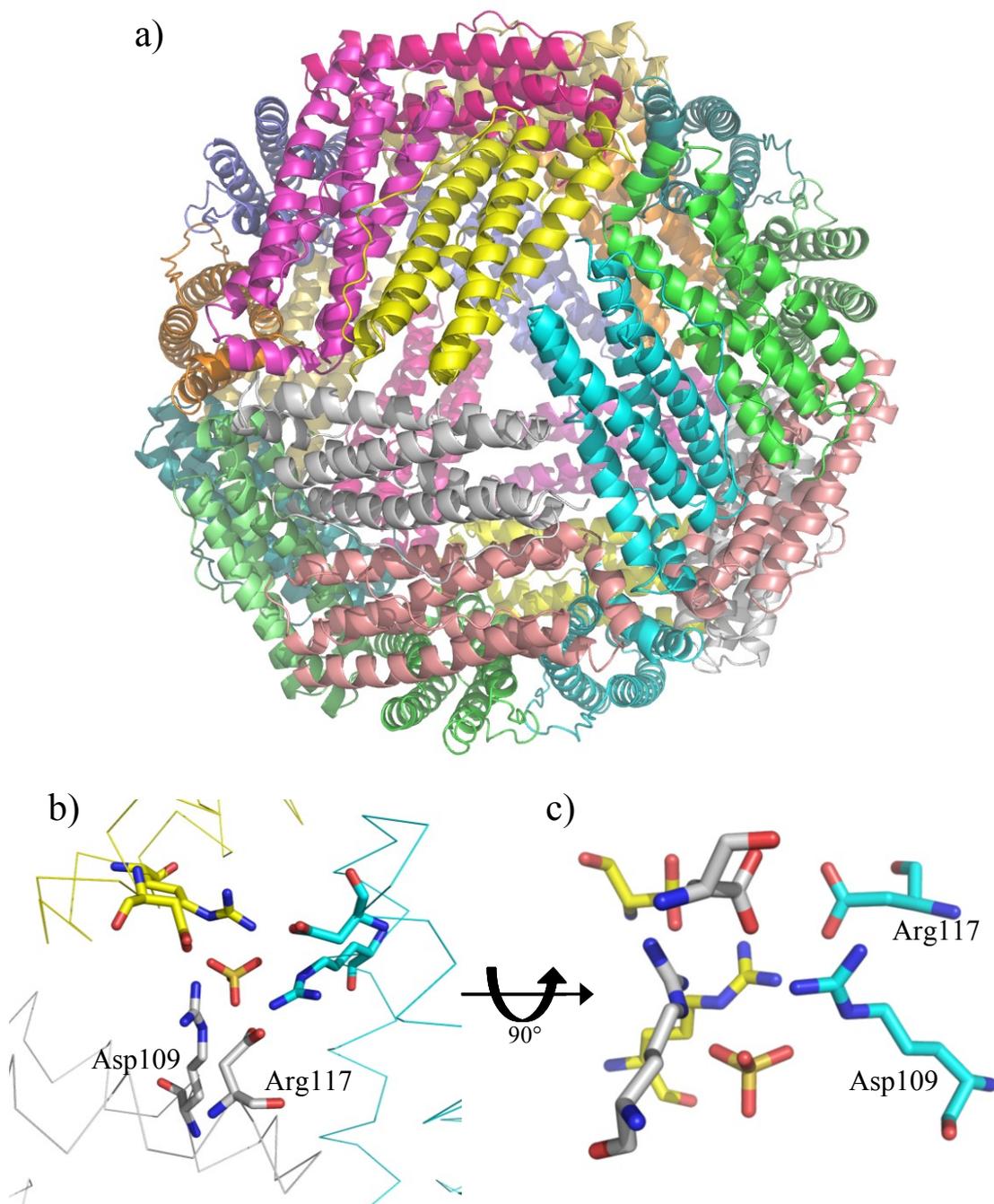
Fig. 7.4. A view along a 3-fold axis of the *E. coli* bacterioferritin. a) Showing the whole bacterioferritin icositetramer. b) Closeup view from a) showing Asp109 and Arg117 from each symmetry related subunit and a sulfate molecule in centre of the channel. c) As in b) with a 90° rotation around the x-axis. (Figure produced using PyMOL, Delano Scientific, USA)
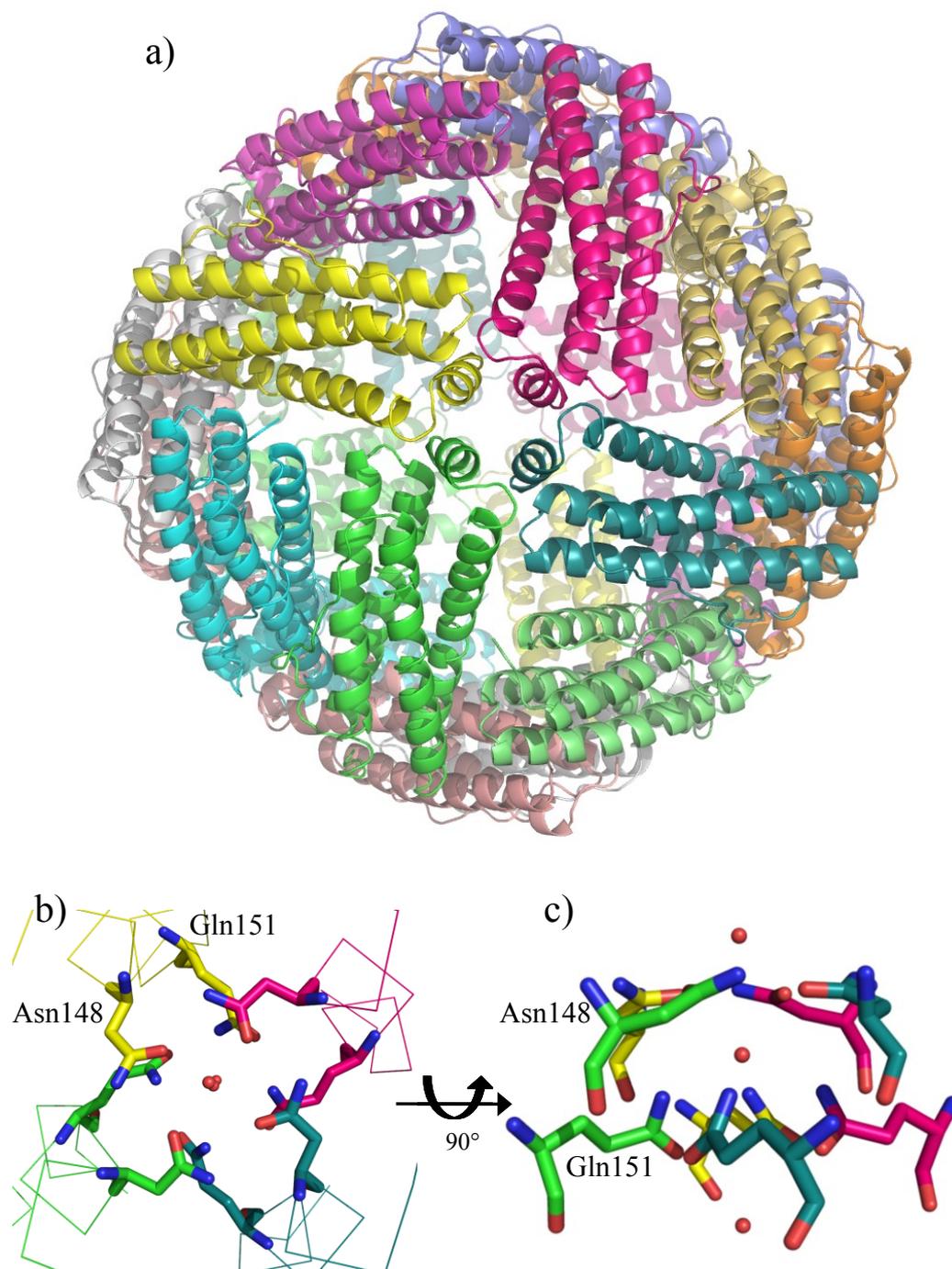
Fig 7.5. A view down a 4-fold axis of the *E. coli* bacterioferritin. a) View down the 4-fold axis showing the whole bacterioferritin icositetramer. b) Closeup view from a) showing Asn148 and Gln151 from each symmetry related subunit and 3 water molecules lined up along the non-crystallographic symmetry axis. c) As in b) with a 90° rotation around the x-axis. (Figure produced using PyMOL, Delano Scientific USA)

7.2.2 Heme Site

The heme molecule in bacterioferritin is bound between two symmetry related monomers on the dimer interface and is chelated by two equivalent Met52 residues, one from each subunit. The heme binding pocket (Fig 7.6) is mainly hydrophobic consisting of Leu19, Ile22, Asn23, Phe26, Tyr45, Ile49, Met52, Lys53, Ala55 and Ile59 from each subunit. The heme is bound in the pocket with its two propionate chains protruding into the internal cavity and is bound in two symmetry related conformations. In the structure presented here the electron density for the heme (Fig 7.7) appears to show no preference for binding in either conformation, in contrast to what was seen for the manganese bound structure (Frolow *et al.*, 1994) where binding in one predominant conformation was observed. The extra density visible in Fig 7.7 can be filled by rotating the heme 180°, thereby placing the vinyl groups of C3 and C8 into the extra observed electron density. In addition the electron density around C8 shows this vinyl group to face in two orientations and has been modelled as such.

The electron density of the heme molecule shows no preference for either of the two conformations observed and is reflected by the symmetrical nature of the interactions between the heme and the residues lining the pocket. An analysis of the atomic contacts made by heme binding in the two alternate conformations (Table 7.1; performed using the WHAT IF web server; Rodriquez *et al.*, 1998) shows there is little difference in the number of contacts made by the two alternate conformations, with heme A making 61 atomic contacts and heme B making 64. This represents less than a 5% difference in the number of contacts made by either possible conformation. The largest differences can be seen with the interaction of Phe26 which makes 9 interactions with heme A, 7 from subunit A and only 2 from subunit B. Likewise a similar situation is observed with heme B having 10 interactions in total, 4 from subunit A and 6 from subunit B. The overall numbers of interactions are similar, but there are differences in the number of interactions contributed by each subunit. This is due to the Phe26 of one subunit interacting with the methyl at C2 of the heme, where the other interacts with the vinyl group at C8 of the heme, producing more interactions. Further small variations in the number of interactions are observed, but these generally represent differences of <0.2Å where one interaction is not recognised since it falls just outside the range to be recognised as an atomic contact. It is

believed that this apparent asymmetry would disappear at higher resolution where more precise measurement of the atomic positions can be made.

Residue	Heme A Subunit A	Heme A Subunit B	No. of contacts with Heme A	Heme B Subunit A	Heme B Subunit B	No. of contacts with Heme B
Leu 19	1	1	2	0	1	1
Ile 22	2	1	3	2	1	3
Asn23	0	0	0	1	0	1
Phe 26	7	2	9	4	6	10
Tyr 45	0	1	1	0	1	1
Ile 49	1	2	3	1	2	3
Met 52	20	21	41	20	21	41
lys 53	1	1	2	1	1	2
Ala 55	0	0	0	1	0	1
Ile 59	0	0	0	1	0	1
Total	32	29	61	31	33	64

Table 7.1. Summary of the atomic contacts made by the two alternate conformations of heme with subunits A and B of *E. coli* bacterioferritin. Inter-atomic contacts calculated using the WHAT IF www servers (Rodriguez *et al*., 1998).
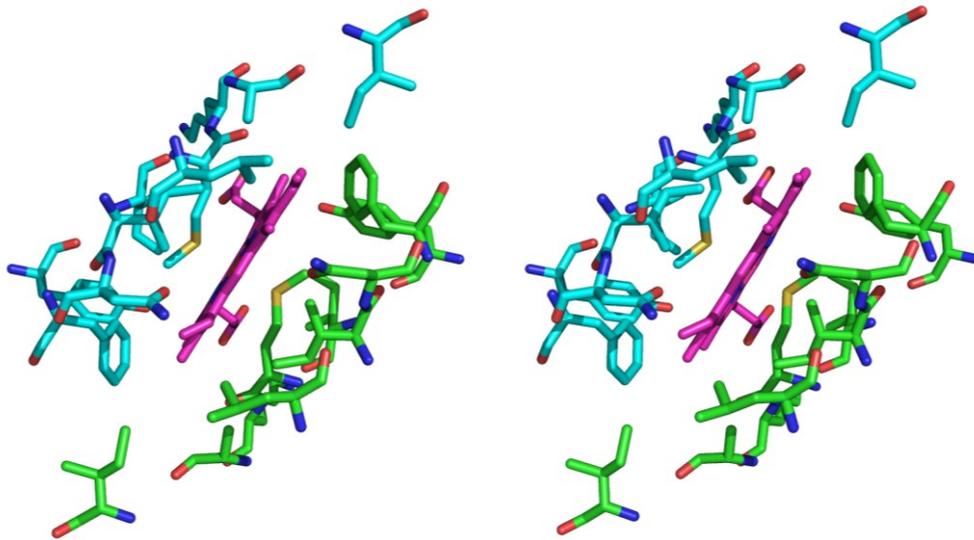
Fig 7.6. A stereo view of the heme in *E. coli* bacterioferritin and all of the interacting residues demonstrating the symmetrical nature of the heme binding site, as viewed from the internal cavity. (Figure produced using PyMOL, Delano Scientific, USA).
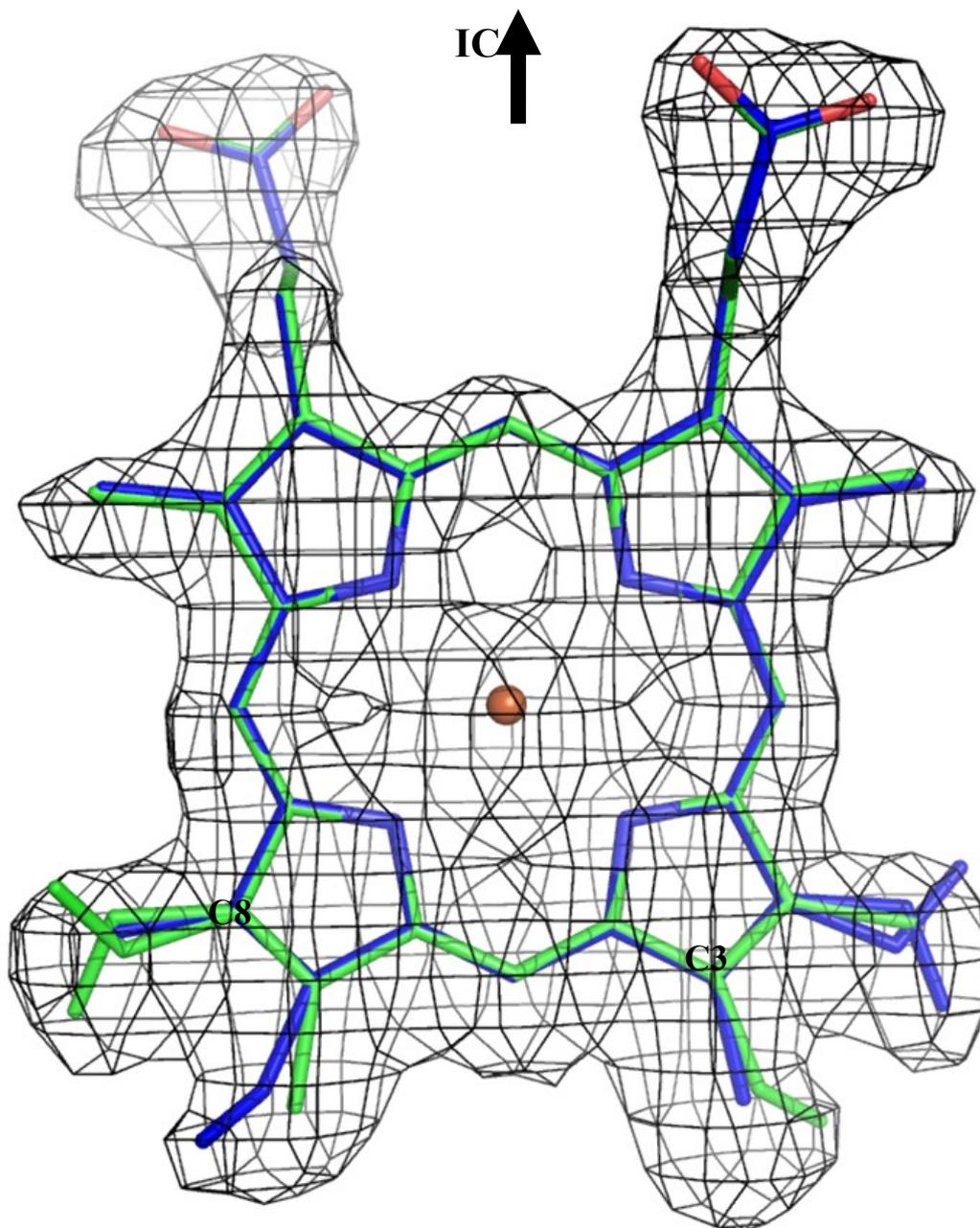
Fig 7.7. $2F_C$-F_0 map of *E. coli* bacterioferritin contoured at 1.5σ, showing one of the two orientations that the heme can adopt. The other orientation is a 180° rotation about the y-axis through the iron. The C8 vinyl also adopts an additional 2nd position, pointing downwards in the diagram. C3 and C8 are labelled for reference on the green molecule. The internal cavity is located at the top of the diagram indicated by arrow (IC). (Picture produced using PyMOL, Delano Scientific, USA).

A water filled pocket is located at the top of the heme binding site and has access to the external medium via two symmetry related pores (Fig 7.8) which pass between loop L and helix A in each subunit. The cavity contains four waters observed in all previous structures of *E. coli* bacterioferritin, but in addition a chain of waters are observed passing from these four waters located deep in the cavity towards the external environment thereby providing a potential link between internal and external environments. This is the only structure to date of all the bacterioferritins to show a direct link between the heme and the external environment. In the *D. desulfuricans* bacterioferritin structures the analogous water cavity is connected to the external environment but is separated from the heme by a pair of symmetry related His28 residues. In the *E. coli* sequence this histidine is conserved, but structurally it is replaced by Asn23. These four waters form an extensive hydrogen bond network within the cavity (Fig 7.9). A pair of Leu19 residues donate one hydrogen bond each from their backbone carbonyls to the lower waters whilst the Asn23 residues form two hydrogen bonds each, one with an upper water and one with a lower water. The lower pair of waters also hydrogen bond with the upper pair of waters which in addition to hydrogen bonding with each other also form hydrogen bonds with waters residing in both symmetry related channels linking them to the external environment. This link between the heme and the external environment may be important for the function of the protein. In horse spleen ferritin loaded with heme an increase in the rate of iron reduction and release from the core was observed due to an increase in the rate of electron transfer to the core mediated by the heme (Kadir *et al*., 1992). The heme in *E. coli* bacterioferritin can therefore interact with the internal and external environments and mediate the transfer of electrons to the internal core and potentially regulate iron release.
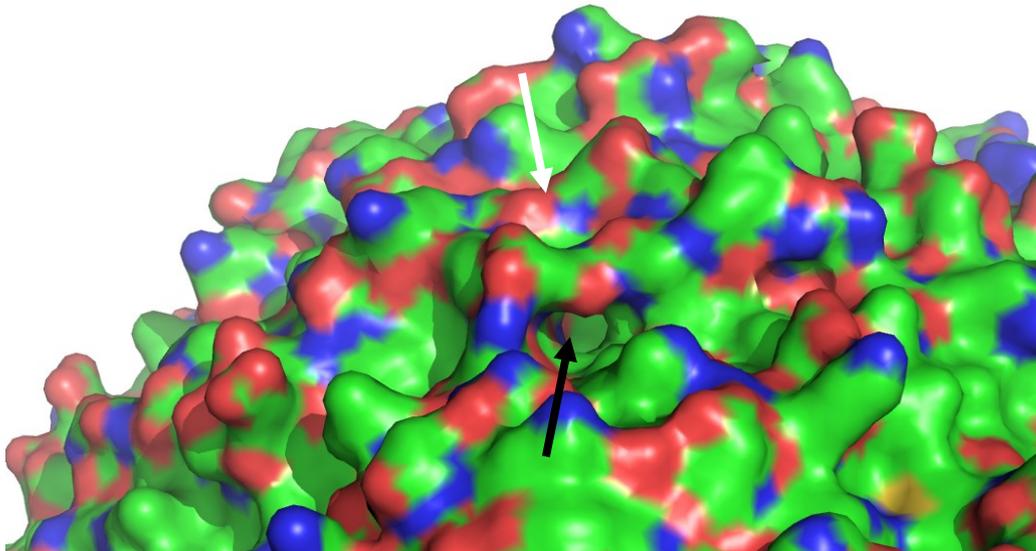
Fig 7.8. Location of the heme water channel in the *E. coli* bacterioferritin is indicated by a black arrow. The symmetrically related pore (not visible) connecting to same heme molecule is indicated by a white arrow. (Figure created using PyMOL, Delano Scientific USA)
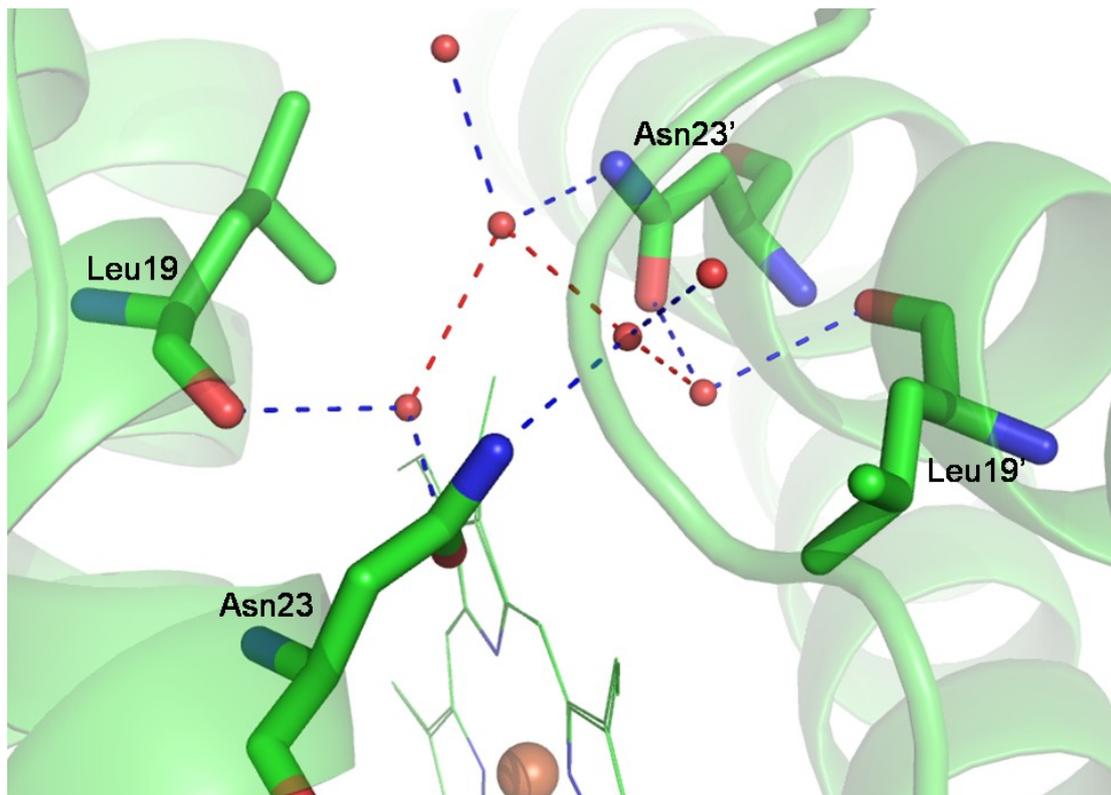
Fig 7.9. The hydrogen bonding network within the heme water cavity of *E. coli* bacterioferritin, showing interacting residues and two additional water molecules residing in the symmetry related pores. Hydrogen bonds are shown as dashed lines. Those connecting the four waters within the cavity are shown in red. (Figure created using PyMOL, Delano Scientific USA)

7.2.3 Ferroxidase Centre

Within the centre of the 4 helix bundle lies a binuclear metal site ligated by residues from helices A, B, C and D. This site usually binds Fe(II) and oxidises it to the insoluble Fe(III) form. The ferroxidase centre has also been shown to bind manganese (Frolow *et al*., 1994), zinc (Le Brun *et al*., 1995) and cobalt (Keech *et al*., 1997). In the structure presented here zinc is seen bound to the ferroxidase centre. The binuclear zinc is coordinated by four glutamates and two histidines (Fig 7.10). His54 from helix B and His130 from helix D both coordinate one zinc each from beneath with their ND1 nitrogens. Glu51 from helix B and Glu127 from helix D both provide axial bridging ligands to both zincs. Glu18 from helix A offers two ligands to the zinc in site 1, whereas the equivalent, Glu94, from helix C offers one ligand for the zinc in site 2. Therefore Zn1 has five protein ligands and a near square pyramidal coordination geometry, Zn2 has four protein ligands and more tetragonal coordination geometry.
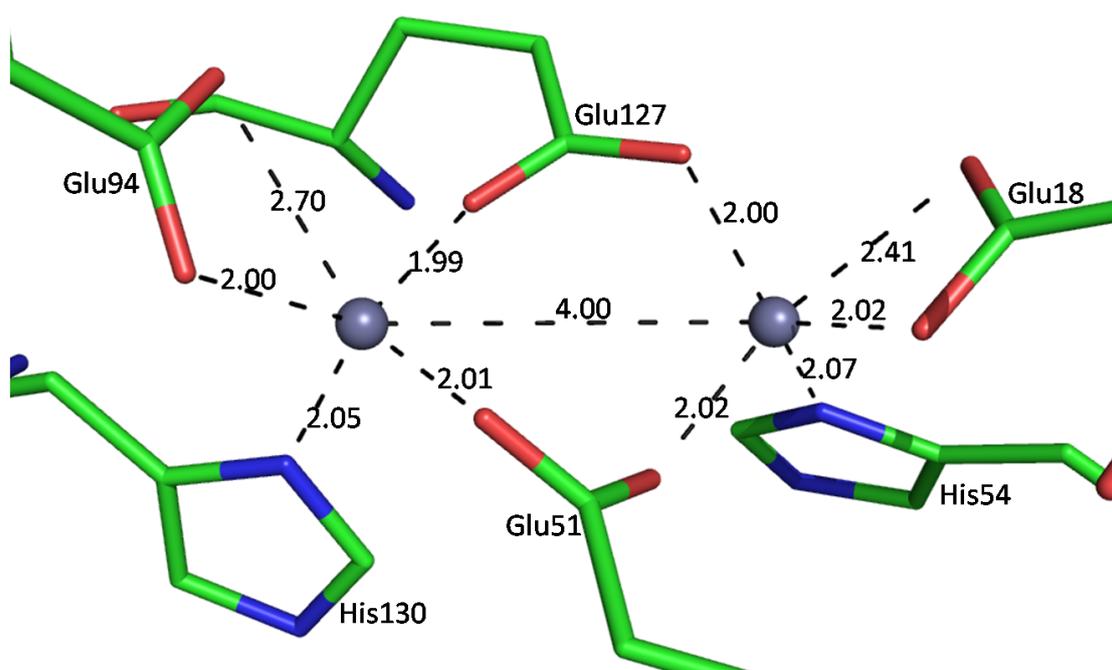
Fig. 7.10. The ferroxidase centre in *E. coli* bacterioferritin, showing the zinc atoms and all of the interacting amino acid residues with interatomic distances shown. (Figure created using PyMOL, Delano Scientific USA)

This difference in the ligands for sites one and two has been seen previously in crystal structures of *E. coli* bacterioferritin (1BFR, 2HTN). The non-equivalence of the metal sites has been demonstrated in kinetic studies where mutants of Glu18 and Glu 94

responded differently to the uptake of metal ions (Keech *et al*., 1997). Partial occupancy of the metal atoms in the ferroxidase site has been observed in previous structures of *E. coli* bacterioferritin (van Eerde *et al*., 2006) and bacterioferritin from other organisms (Liu *et al*., 2004; Swartz *et al*., 2006; Macedo *et al*., 2003). The B-factors of the zinc atoms show only a very small difference suggesting that both metal atoms are of a similar occupancy. The Zn1 has an average B-factor of 21.27$Å^3$ (21.34-21.75) where as Zn2 has an average B-factor of 24.19$Å^3$ (23.76-24.65). In addition the OE1 and OE2 of Glu51 and Glu127 also display a similar asymmetry in B-factors. Zn2 may have a slightly lower occupancy to that of Zn1, but this has not been reflected in the final model since the B-factors are reasonably close and any reduction in occupancy is only likely to be marginal.

7.2.4 Iron Entry and Exit Paths

There are many potential routes for iron entry and exit from the core. The bacterioferritins, bacterial ferritins, and mammalian ferritins all vary dramatically at these pores potentially leading to different mechanisms of core formation. The bacterioferritins from different organisms also vary amongst themselves and may use these pores in different ways.

Situated directly above the ferroxidase site are a large cavity and a channel linking to the outside environment (Fig 7.11). This cavity is approximately 16Å long in the Zn1-Zn2 direction, and 9Å across in the zinc-heme direction and encloses on average 9 waters (8-10). This cavity and its entrance are highly negatively charged and would therefore be able to attract cations such as Fe^{2+} towards the ferroxidase centre. It has been suggested that through a concerted movement of His130 and release of at least one iron from the ferroxidase site would open up a clear channel allowing access to the core (Macedo *et al*., 2003). Partial occupancy of site 2 in the manganese and mixed iron/zinc structures suggest this may occur but neither have shown movement of His130 and opening up of a clear channel. In the zinc bound form presented here both sites appear to be fully occupied, which may not be surprising given the high affinity of zinc for the ferroxidase site (Keech *et al*., 1997).
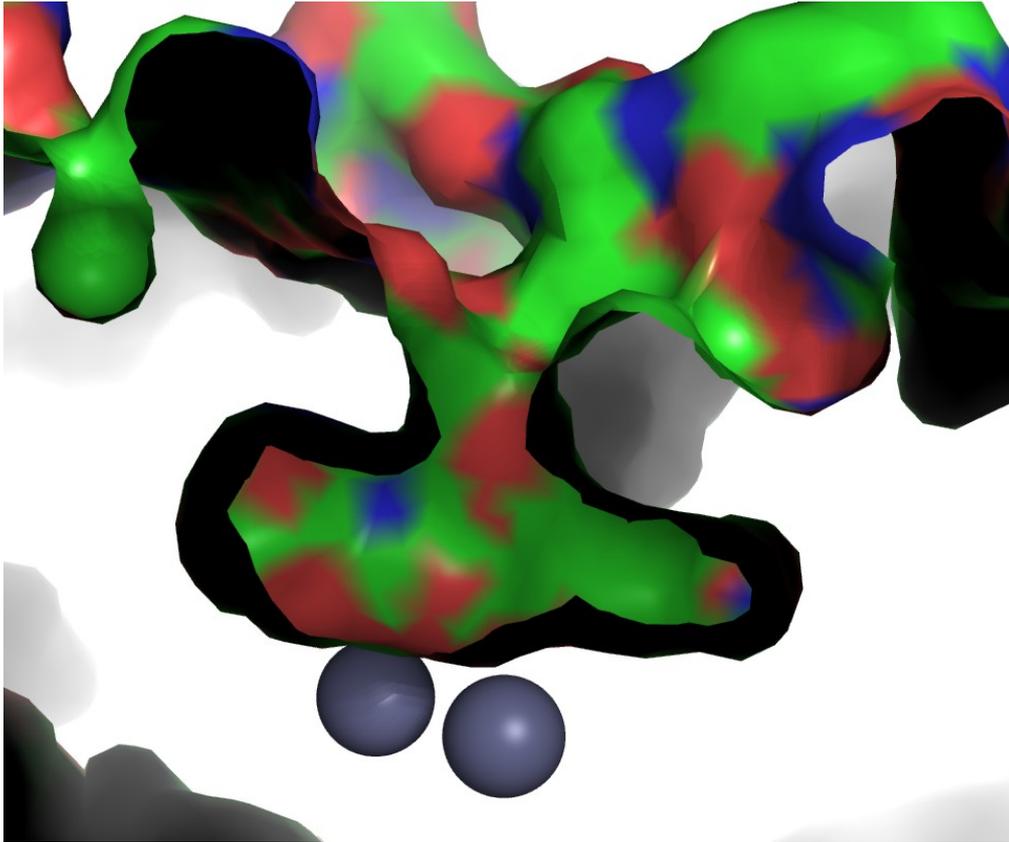
Fig 7.11.   The ferroxidase cavity and channel of *E. coli* bacterioferritin shown as a surface representation in cross section.   Zinc atoms are shown in grey.   (Figure created using PyMOL, Delano Scientific USA)

Located at the 3-fold and 4-fold symmetry axis are subunit interfaces where it has been previously suggested (Dautant *et al*., 1998) that channels exist and provide a route for Fe(II) to pass from the external environment to the core and vice versa. The 3-fold channel (Fig 7.4) consists of three Asp109 residues on the external surface and three Arg117 residues from symmetry related subunits lying underneath these with access to the internal solvent. Further into the centre is a sulfate interacting with the three arginine residues and a large cavity containing 16 water molecules.   At the internal entrance to this cavity (Fig 7.12) are three Asp118 and three Glu122 residues providing a highly negatively charged environment.
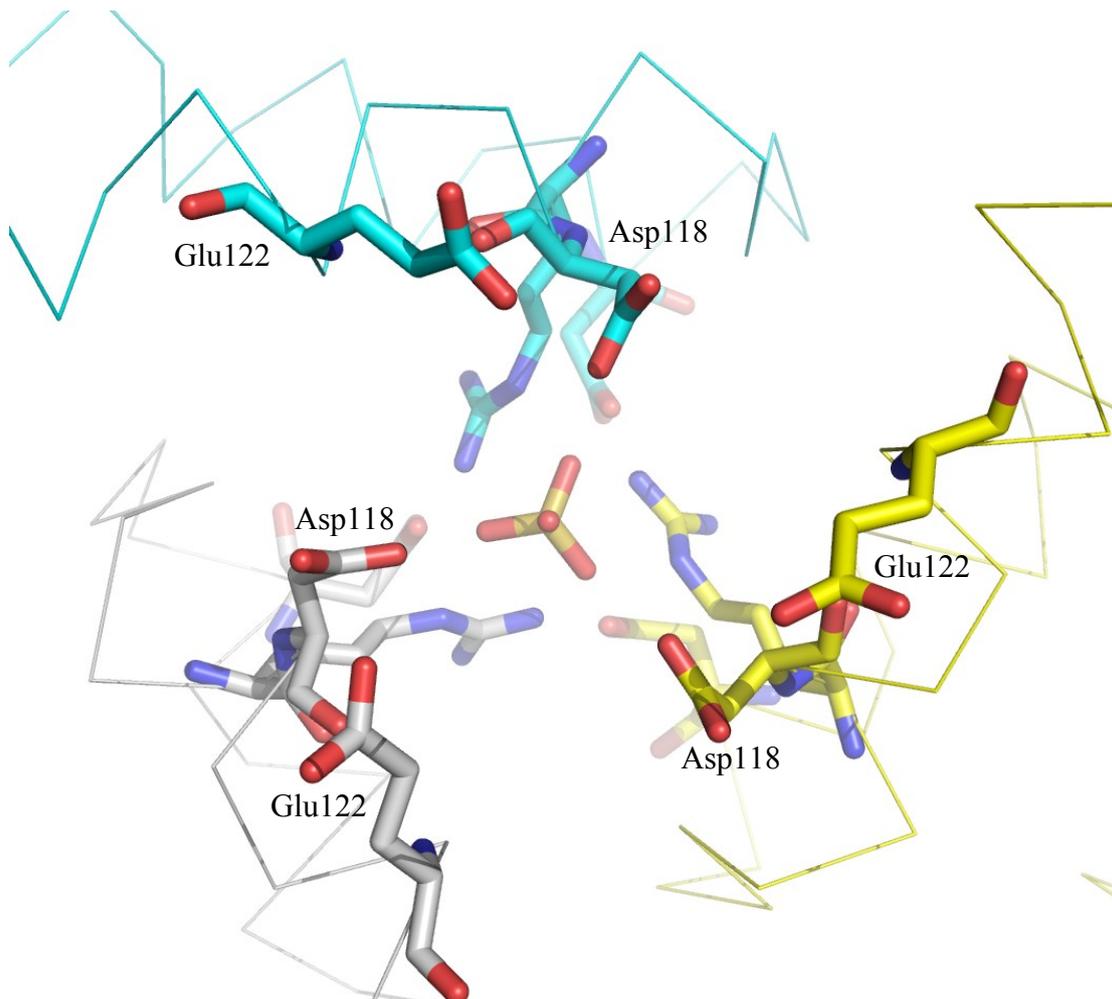
Fig 7.12, View of a 3-fold axis channel of *E. coli* bacterioferritin from the internal cavity. The figure shows Asp109 and Arg117 (background, not labelled), and Asp118 and Glu122 (foreground, labelled) from 3 symmetry related subunits, and a central sulfate ion bound at the top of the channel.

The 4-fold axis is at the point where four short E helices meet (Fig 7.5) with the channel entrance consisting of Asn148 and Gln151 from four symmetry related subunits with Asn148 lying on the outer surface and Gln151 lying underneath contributing to the inner surface. Manganese has been observed to be bound in the centre of this channel by Gln151s (Dautant *et al.*, 1998). Metal atoms were also found within this channel in the *A. vinlandii* bacterioferritin structures (Swartz *et al.*, 2006; Liu *et al.*, 2004). No metal atoms were observed in this structure at this position, presumably due to the absence of any heavy metals in the crystallisation solutions. This position between the glutamines and asparagines is occupied by a single water molecule 3.8Å from a water located on the outer surface and 5.5Å from a

water located on the inner surface within a cavity containing 11 waters. In comparison to the 3-fold channel the surfaces of the 4-fold channel are relatively uncharged (Carrondo, 2003). The 3-fold and 4-fold channels in ferritins and bacterioferritins from other organisms vary considerably. The 4-fold axis in mammalian L chain ferritins is highly hydrophobic, whilst mammalian H chain ferritins have some histidine and methionine residues compared with the hydrophilic channels in *E. coli* bacterioferritin. The 3-fold channel of *D. desulfuricans* bacterioferritin is highly positively charged on the external surface, the opposite to what is observed for the *E. coli* bacterioferritin.

Another potential channel is located where a dimer meets another, head to side on (Fig 7.14a), and have been termed B-channels. These channels were identified in *D. desulfuricans* bacterioferritin (Macedo *et al*., 2003) due to the large concentration of negatively charged residues around the channel. The B-channel (Fig 7.14b) in *E. coli* bacterioferritin consists of Met1, Phe64 and Glu66 from one subunit, Asn34 and Trp35 from the second, making a dimer with the first and Asp132, Glu135, Thr136, and Asp139 from the third subunit. This channel provides the most easily accessible route of entry for iron directly to the core and consisting of predominantly negatively charged amino acids would strongly attract cations.
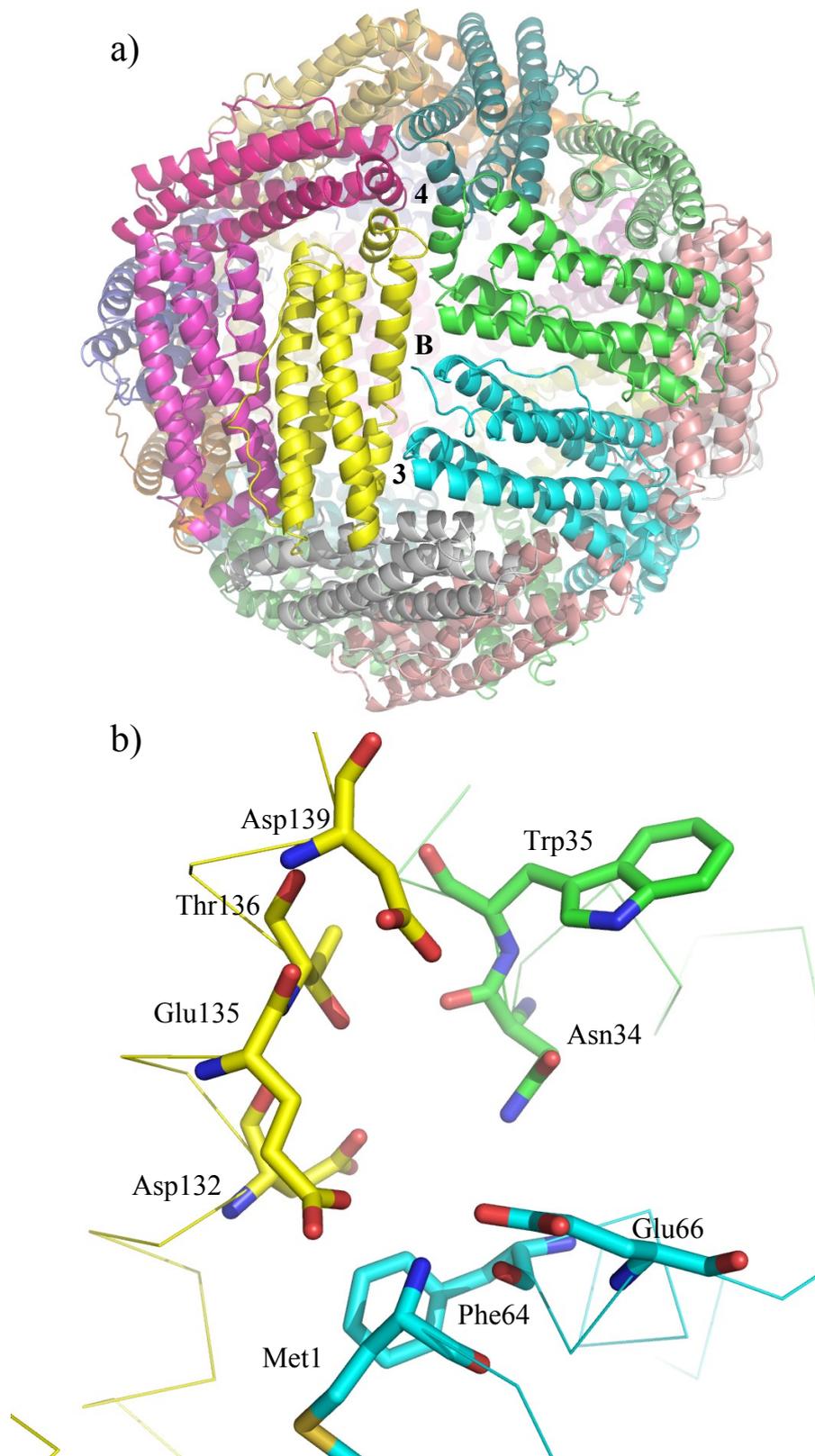
Fig 7.13. An external view of the B-channel in *E. coli* bacterioferritin. a) showing the whole icositetramer with the 4-fold (4), 3-fold (3) and B-channel (B) labelled. b) view of all the amino acids surrounding the channel.

Whilst it has been postulated that there are a variety channels linking the internal and external environments, it has not been shown which channels are capable of allowing Fe(II) ions to pass through in the *E. coli* bacterioferritin. The ferroxidase channel is very likely to supply the ferroxidase site with Fe^{2+} ions, the mechanistic fate of which is still rather unclear. The 3-fold, 4-fold and B-channels in *E. coli* bacterioferritin have a similar amino acid composition to those in the *A. vinlandii* bacterioferritin, the structures of which have shown metals bound at all of these channels (Swartz *et al*., 2006; Liu et al., 2004), the metals bound at the 4-fold channel are bound in a similar fashion to the manganese found in previous *E. coli* bacterioferritin structures (Frolow *et al*., 1994) suggesting the 4-fold channel acts as a route for iron entry and/or exit in these bacterioferritins. The *A. vinlandii* bacterioferritin structures also provide clues to the function of the 3-fold channels. Rather than being a route for iron entry it has been suggested (Liu *et al*., 2004) that during core formation the bottom of the channel may become positively charged and build up a large electrostatic gradient allowing phosphate ions to flow into the core. In the *E. coli* bacterioferritin structure described here, a sulfate molecule was found located in the middle of the 3-fold channel. Sulfate is of a similar size and co-ordination geometry to phosphate, although it has +2 formal charge rather than +3 of phosphate, suggesting phosphate may be able to occupy a similar position en-route to the core. The B-channel possesses many negatively charged residues and provides a good candidate for iron entry or exit. However this channel lacks much of the specificity shown by the other channels, and its specific function is yet to be established.

8 Structural basis of AKR substrate specificity

8.1 Cofactor binding

All the members of the AKR superfamily with three dimensional structures bind the NADPH cofactor in the same orientation, ready to donate the 4-*pro-R* hydride from the cofactor nicotinamide. The only exception to this is the AKR11A1 (Ehrensberger and Wilson, 2004) which binds cofactor with the ribose and the nicotinamide rotated by 180° relative to the other AKR family members (Fig 8.1) which positions the 4-*pro-S* hydride ready for transfer to the substrate, and has reduced catalytic activity.
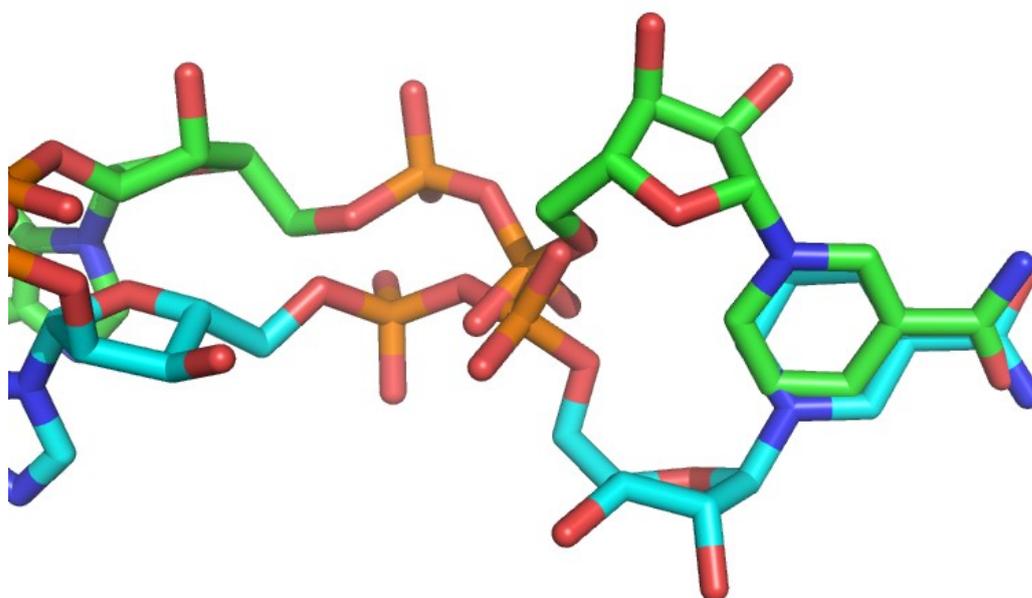
Fig 8.1. Figure showing the alternate binding conformation of NADPH in AKR11A1 (green), with the ribose and nicotinamide moiety rotated by 180° relative to the cofactor in AKR11B1 (cyan). (Figure produced with PyMol, Delano Scientific, USA)

The TmAKR cofactor binding site is devoid of NADPH and instead binds two sulphate molecules from the crystallisation medium. These occupy positions analogous to the phosphates of the NADPH cofactor (Fig 8.2), and are bound by similar residues which bind the cofactor phosphates in other AKR family members. The cofactor binding site is highly conserved in sequence and structure (Jez et al.,

1997) to such an extent that we are able to superimpose the active site from AKR11B1 and its NADPH cofactor onto the TmAKR structure (Fig 8.3). The superimposed cofactor sits in a clearly defined cofactor binding pocket, and interacts with residues conserved across the AKR superfamily (See Chapter 6).
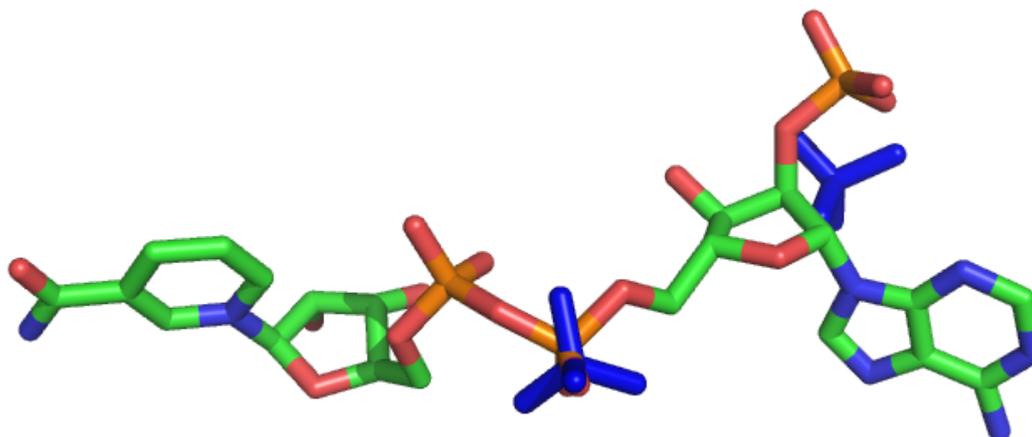
Fig 8.2. Superimposition of NADPH from AKR11B1 (1PZ1) and the two sulphates (shown in blue) bound in the cofactor binding site of TmAKR. (Figure produced with PyMol, Delano Scientific, USA)
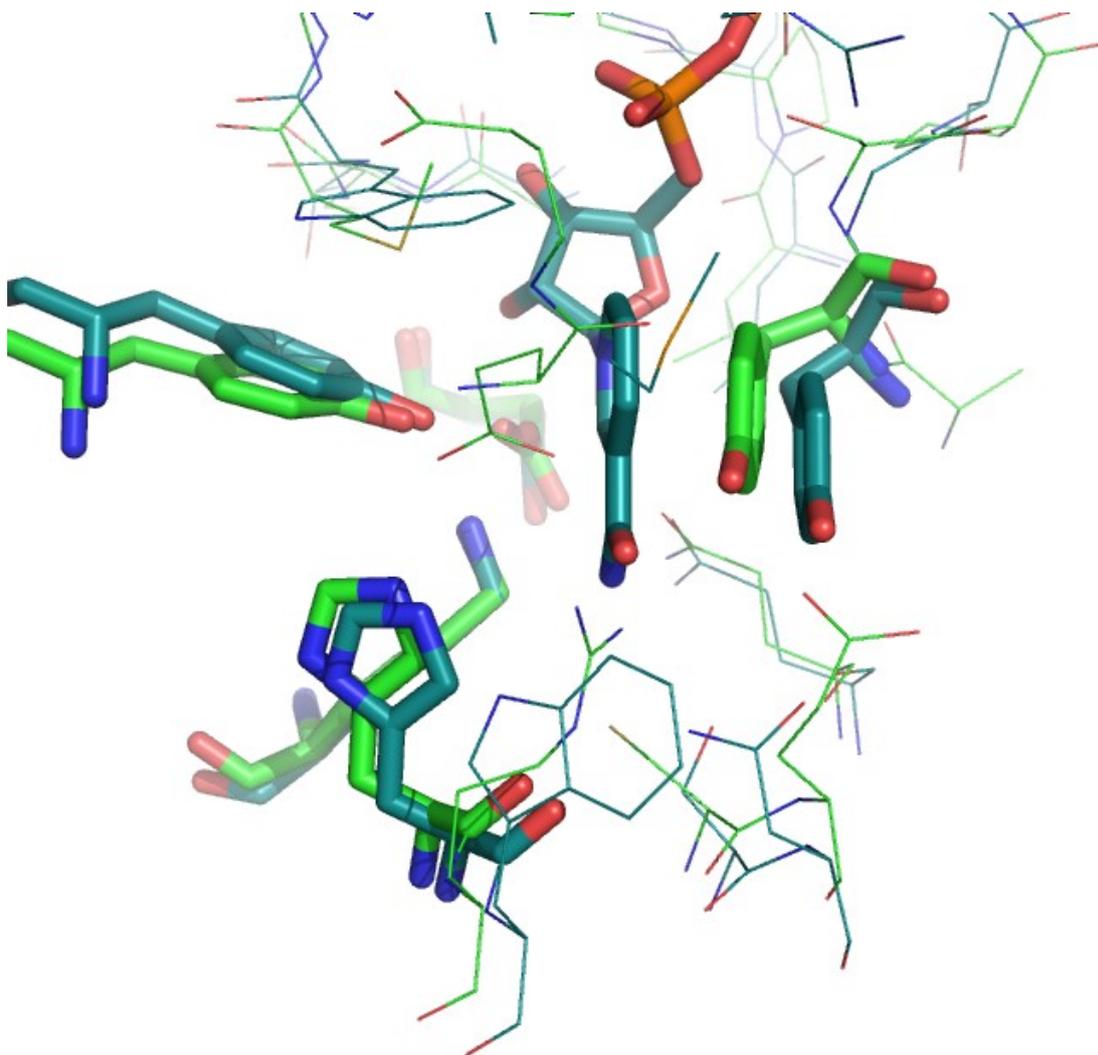
Fig 8.3. Superimposition of TmAKR with AKR11B1 showing NADPH cofactor binding site and its structural conservation. The catalytic tetrad and stacking tyrosine and shown as sticks, other residues neighbouring NADPH are shown as lines. (Figure produced with PyMol, Delano Scientific, USA)

8.2 Loop structures

The three dimensional structure of the AKRs is highly conserved across all superfamily members, particularly the central core composed of the $\alpha_8\beta_8$ barrel. Most of the variability across the AKRs is present within the loop regions (See section 6.1.1 for nomenclature scheme) and the C-terminal loop. These variable portions of the AKRs are responsible for the substrate specificity of the enzyme (Jez et al., 1997).

A sequence alignment of AKRs structurally related to the TmAKR enzyme (Appendix 8) shows the conservation of the core structure and the variability present in the loop regions. This variability in loop sequence and size correlates with the substrate specificity of the AKRs. The most obvious difference in loop structure occurs on loops 3 and 4 (Fig 8.4), where loop 4 of AKR1A1 occupies the same three dimensional space as loop 3 of TmAKR. TmAKR and the AKR11 family members all possess a longer loop 3 and shorter loop 4 than most other AKRs, TmAKR also has a longer Loop 6 and shorter C-terminal loop. The loop 4 has been implicated to have an important role in substrate binding (Jez et al., 1997), and alteration of the sequence and/or size of this loop may be expected to have an impact upon substrate binding. In contrast to TmAKR, AKR1A2 has the largest loop 4 of those investigated and binds a much larger substrate than TmAKR (C15 compared with C8). TmAKR along with the related AKR11A and B (not 11C) and the *E. coli* TAS protein all have a longer loop 6 than the other AKRs in appendix 8 along with a shorter C-terminal loop. The C-terminal loop also has an important role in substrate binding, it maybe that with a reduction in the C-terminal loop and loop 4, some of the substrate specificity has been transferred to loops 6 and 3 respectively and thereby producing an alteration in the overall substrate specificity.
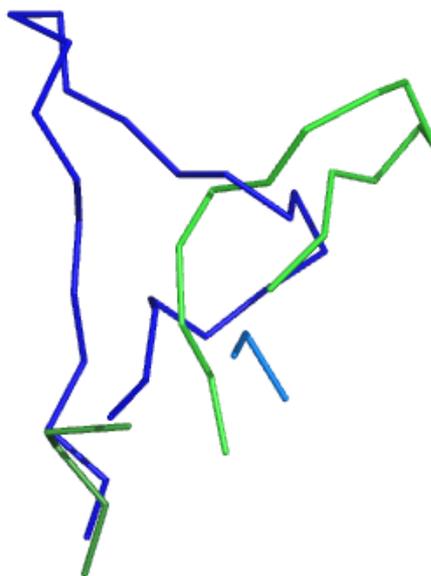
Fig 8.4. Variation in Loops 3 and 4 of the AKRs. TmAKR shown in greens, with loop 3 in light green and loop 4 in dark green. AKR1A1 shown in blues with loop 3 in light blue and loop 4 in dark blue. (Figure produced with PyMol, Delano Scientific, USA)

8.3 Substrate binding

The binding of cofactor to the active site partially assembles the substrate binding site, and it becomes fully formed upon substrate binding as the substrate binding loops enclose around the substrate (Sanli & Blaber 2001, Sotriffer et al., 2004, Couture et al., 2005). The substrate binding loops make a significant contribution towards the substrate binding site, and since these are missing in the TmAKR structures it is hard to predict which loop residues interact with the substrate. In hydroxysteroid dehydrogenases the steroid substrate is either bound pointing directly upwards away from the active site or lying across the top of the catalytic tyrosine (Fig 8.5). In the TmAKR structure there is space for the phenyl ring of the substrate to bind near to the catalytic tyrosine and Phe87/Trp97 (Site B in Fig 8.6), a space analogous to epitestosterone binding shown in Fig 8.5.
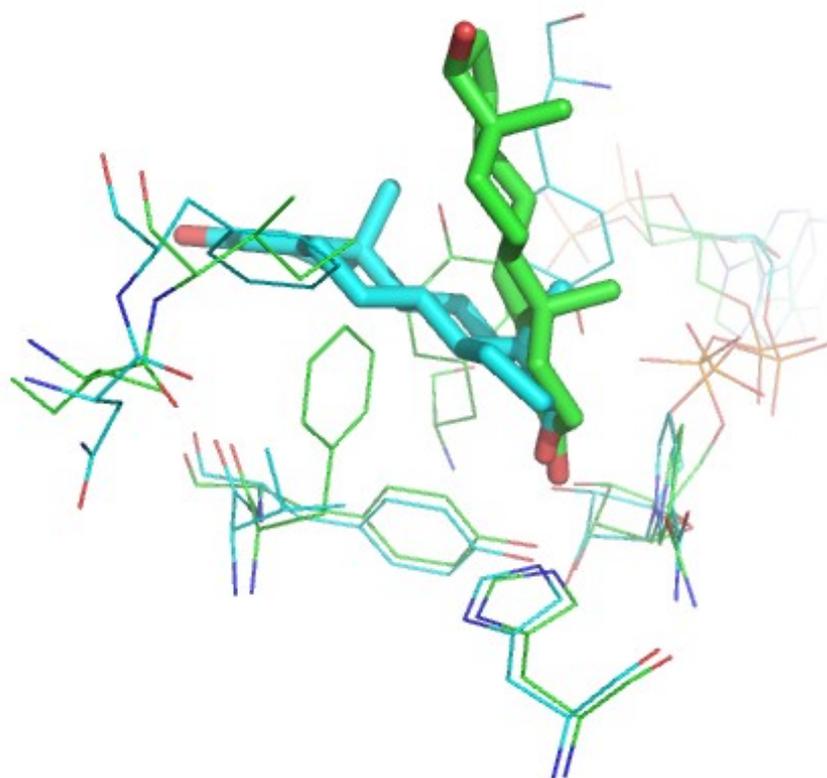
Fig 8.5. Steroid substrate binding in the hydroxysteroid dehydrogenases. Testosterone binding in 1q13 shown in green, and epi-testosterone binding in 2ipf shown in cyan. (Figure produced with PyMol, Delano Scientific, USA)

The substrate specificity of TmAKR has been shown to be specific for aldehydes and not ketones (Chapter 3). This presents the prospect of engineering the substrate specificity to accept ketones, allowing for the production of secondary alcohols. There are many methods available to engineer an enzymes activity, these can broadly be separated into rational design and directed evolution. Rational design requires knowledge of the three dimensional structure of the protein of interest either in the form of an X-ray or NMR structure, or a homology model, we also need to be able to make accurate predictions of the mutation effects which is not always easy. In the case of TmAKR we lack information on the mode of substrate binding, which decreases the chances of success with rational design.

Directed evolution can be a very efficient method to engineer an enzymes properties. However an efficient screening/selection method is required to allow easy identification of positive mutants (Bershtein & Tawfik, 2008). Selection methods rely on the presence of a relevant auxotrophic strain to allow genetic complementation

with positive mutants (Boersma *et. al.,* 2007), but allows larger library sizes to be screened (up to 10^{13} variants) compared to selection methods ($\approx 10^4$ variants) (Cipolla 2004). If screening is to be employed an efficient assay is required to allow analysis of every mutant, and is often the rate limiting step in the evolution experiment.

Error prone PCR (epPCR) is often performed as an initial step during directed evolution. epPCR uses non-optimal conditions to amplify the gene of interest and introduce mutations at a set frequency. The drawback of epPCR is the screening effort which is required, for an enzyme the size of TmAKR (333 aa's) with a frequency of one amino acid change per mutant, there are 6327 possible mutants, if the frequency is increased to two or three mutations then the possible number of mutants rises to 16 million or 30 billion respectively (Arnold & Georgiou, 2003), representing an unrealistic screening effort. In reality however it is not possible to generate all these mutants, due to the bias introduced from the genetic code (Wong *et al.,* 2006). Despite these drawbacks, epPCR is an effective method for directed evolution as it allows exploration of sequence space across the whole gene and identification of hot spots which can be exploited in further rounds of directed evolution.

Saturation mutagenesis may be carried out at these identified hot spots or at predetermined sites from the structural information available. Combinatorial active-site saturation testing (CASTing) randomises two or three amino acids adjacent to the active site in combination (Clouthier *et al.,* 2006, Reetz *et al.,* 2007). In this way we are able to identify mutations which are additive in nature and combinations which would otherwise be missed from epPCR. Using a CASTing approach the screening effort is greatly reduced since only two or three amino acids are focused upon rather than the whole protein. Iterative saturation mutagenesis (ISM) builds upon the concept of CASTing (Reetz and Carballeira, 2007). Sites are identified in much the same way as for CASTing and subjected to saturation mutagenesis; the method varies by identifying multiple sites and creating a library individually for each site. The most successful mutant from the first round across all sites/libraries is identified for use in the next round. During further rounds the remaining sites are again subjected to saturation mutagenesis, allowing for identification of new mutants with additive effects. The effectiveness of this method has been demonstrated by the case of *Aspergillus niger* epoxide hydrolase where the enantioselectivity was increased 25-

fold from screening 20,000 clones (Reetz et al., 2006), where as screening a similar number of clones from epPCR produced a 2.5-fold enhancement (Reetz, et al., 2005). CASTing and ISM have the potential to enable the engineering of the TmAKR activity to allow it to accept ketone substrates and also produce an enzyme with high enantioselectivity (Reetz, 2004). Sites within the TmAKR active site have been identified (Fig 8.6) which could be the subject of an ISM experiment. Glutamate 227 occupies a position within the active site analogous to the binding site of the substrate carbonyl and has been used as the centre of the active site for the purpose of identifying relevant ISM sites. Target residues have been identified based on proximity to Glu227, and grouped into seven sites (Table 8.1). Those residues which are part of the catalytic tetrad or have an important role in cofactor binding have been excluded from these groups as they would be expected to reduce or remove all catalytic activity.
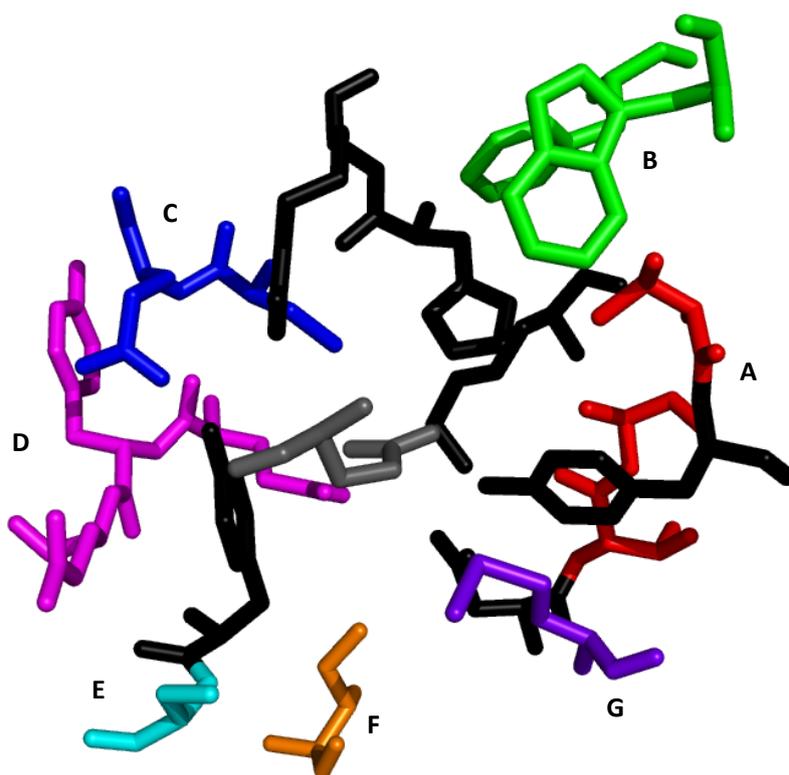
Fig 8.6. Sites identified for iterative saturation mutagenesis (ISM). Essential catalytic and cofactor binding residues excluded from ISM are shown in black. Glu227 occupying the site of substrate binding shown in grey. (Figure produced with PyMol, Delano Scientific, USA)

Site	A	B	C	D	E	F	G
Residues	Thr55	Phe87	Cys159	Gln179	Ser208	Ile281	Met23
	Ala56	Trp97	Gln160	Tyr180			
	Val58			Glu181			

Table 8.1. Summary of sites identified for iterative saturation mutagenesis.

Possibly the best approach to take with the directed evolution of TmAKR to achieve activity towards ketones may be a combination approach of epPCR followed by ISM. Despite the drawbacks of epPCR it is often used effectively as a first step in a directed evolution process. This is because residues are sampled across the whole sequence and may identify residues not considered during ISM. The success of epPCR can also be attributed to long range effects (Reetz, 2002 & Bocola *et al.,* 2004), which may produce effects which may not be predictable. In this way epPCR may be able to identify mutations in TmAKR which are able to introduce activity towards ketones. Once mutations which are capable of altering the activity towards ketones have been identified, then ISM may be used to further explore these newly identified sites and those previously identified around the active site (Fig 8.6). In this way the desired activity may be induced during epPCR, followed by optimisation of the activity and chiral specificity enabling the producing of two evolved enzymes with opposite enantioselectivities.

References

Aitken-Rogers H., Singleton C., Lewin A., Taylor-Gee A., Moore G.R., Le Brun N.E. (2004) Effect of phosphate on bacterioferritin-catalysed iron(II) oxidation. *J. Biol. Inorg. Chem.* **9**, p161-170

Altschul S.F., Gish W,. Miller W., Myers E.W., Lipman D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, p403-410

Andrews S.C., Robinson A.K., Rodriguez-Quinones F. (2003) Bacterial iron homeostasis. *FEMS Microbiology reviews* **27**, p215-237

Anderson S., Marks C.B., Lazarus R., Miller J., Stafford K., Seymour J., Light D., Rastetter W., Estell D. (1985) Production of 2-keto-L-gulonate, an intermediate in L-ascorbate synthesis, by a genetically modified *Erwinia herbicola*. *Science* **230**, p144-149

Arnold F.H., Georgiou G. (eds.) *Directed enzyme evolution: screening and selection methods* (Humana Press, Totowa, New Jersey, 2003)

Baaghil S., Thomson A.J., Moore G.R., Le Brun N.E. (2002) Studies of copper(II)-binding to bacterioferritin and its effects on iron(II) oxidation. *J. Chem. Soc., Dalton Trans.* p811-818

Baaghil S., Lewin A., Moore G.R., Le Brun N.E. (2003) Core formation in *E. coli* bacterioferritin requires a functional ferroxidase centre. *Biochemistry* **42**, p14047-14056

Bershtein S., Tawfik D.S. (2008) Advances in laboratory evolution of enzymes. *Current Opinion in Chemical Biology* **12**, p151-158

Blochl E., Rachel R., Burggraf S., Hafenbradl D., Jannasch H.W., Stetter K.O. (1997) *Pyrolobus funarii*, gen. and sp. nov., represents a novel group of archaea, extending the upper temperature limit for life to 113 degrees C. *Extremophiles* **1**, p14-21

Blow D. (2002) *Outline of crystallography for biologists*. Oxford University Press

Bocola M., Otte N., Jaeger K., Reetz M.T., Thiel W. (2004) Learning from Directed Evolution: Theoretical Investigations into Cooperative Mutations in Lipase Enantioselectivity. *ChemBioChem.* **5**, p214-233

Boersma Y.L., Droge M.J., Quax W.J. (2007) Selection strategies for improved biocatalysts. *FEBS Journal* **274**. p2181–2195

Bohren K.M., Grimshaw C.E., Lai C.J., Harrison D.H., Ringe D., Petsko G.A., Gabbay K.H. (1994) Tyrosine-48 is the proton donor and histidine-110 directs substrate stereochemical selectivity in the reduction reaction of human aldose reductase: enzyme kinetics and crystal structure of the Y48H mutant enzyme. *Biochemistry* **33**, p2021-2032

Bond C.S. (2003) Topdraw: a sketch pad for protein structure topology cartoons. *Bioinformatics* **19**, p311-312

Bonisch H., Backmann J., Kath T., Naumann D., Schafer G. (1996) Adenylate kinase from Sulfolobus acidocaldarius: expression in Esherichia coli and characterisation by fourier transform infrared spectroscopy. *Arch. Biochem. Biophys.* **333**, p75-84

Bou-Abdallah F., Lewin A.C., Le Brun N.E., Moore G.R., Chasteen N.D. (2002) Iron detoxification properties of *Escherichia coli* bacterioferritin. *J. Biol. Chem.* **40**, p37064-37069

Cacciapouti G., Porcelli M., Bertoldo C., De Rosa M., Zappia V. (1994) Purification and characterization of extremely thermophilic and thermostable 5'-methylthioadenosine phosphorylase from the archaeon *Sulfolobus solfataricus*. Purine

nucleoside phosphorylase activity and evidence for intersubunit disulfide bonds. *J. Biol. Chem.* **269**, p24762-24769

Campbell J.L., Hopman T.L., Maxwell J.A., Nejedly Z. (2000) Chemical state sensitivity by PIXE, using solid state detectors. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* **170**, p193-204

Canaves J.M., Page R., Wilson I.A., Stevens R.C. (2004) Protein biophysical properties that correlate with crystallization success in *Thermotoga maritima*: maximum clustering strategy for structural genomics. *J. Mol. Biol.* **344**, p977-991

Cary S.C., Shank T., Stein J. (1998) Worms bask in extreme temperatures. *Nature* **391**, p545-546

Cianci M, Antonyuk S, Bliss N, Bailey MW, Buffey SG, Cheung KC, Clarke JA, Derbyshire GE, Ellis MJ, Enderby MJ, Grant AF, Holbourn MP, Laundy D, Nave C, Ryder R, Stephenson P, Helliwell JR, Hasnain SS (2005) A high-throughput structural biology/proteomics beamline at the srs on a new multipole wiggler. *Journal of Synchrotron Radiation* **12**, p455-466

Cipolla L. (2004) Combinatorial libraries of biocatalysts: application and screening. *Comb Chem High Throughput Screen* **7**, p101–114.

Clark A.T., McCrary B.S., Edmondson S.P., Shriver J.W. (2004) Thermodynamics of core hydrophobicity and packing in the hyperthermophile proteins Sac7d and Sso7d. *Biochemistry* **43**, p2840-2853

Clouthier C.M., Kayser M.M., Reetz M.T. (2006) Designing New Baeyer−Villiger Monooxygenases Using Restricted CASTing. *J. Org. Chem.* **71**, p8431-8437

Collaborative Computational Project, Number 4 (1994). The CCP4 suite: programs for protein crystallography. *Acta. Cryst.* **D50**, p760-763

Connelly S. (2006) Studies on pyroglutamyl carboxypeptidase enzymes and cholinesterase inhibitors: implications for Alzheimer's diseases. Ph.D. Thesis, University of Exeter, Exeter, U.K.

Conway T., Sewell G.W., Osman Y.A., Ingram L.O. (1987). Cloning and sequenceing of the alcohol dehydrogenase II gene from *Zymomonas mobilis*. *J. Bacteriol.* **169**, p2591-2597

Couture J-F., De Jesus-Tran K.P., Roy A-M., Catin L., Cote P-L., Legrand P., Luu-The V., Labrie F., Breton R. (2005) Comparison of crystal structures of human type 3 3α-hydroxysteroid dehydrogenase reveals an "induced-fit" mechanism and a conserved basic motif involved in binding androgen. *Protein Science* **14**, p1485-1497

Cowtan K (1994) Jnt CCP4/ESF-EACBM Newsl. *Protein. Crystallogr.* **31**, p34-38

Daniel R.M., Danson M.J., Eisenthal R., Lee C.K., Peterson M.E. (2007) New parameters controlling the effect of temperature on enzyme activity. *Biochem. Soc. Trans.* **35**, p1543-1546

D'Arcy A., Elmore C., Stihle M., Johnston J.E. (1996) A novel approach to crystallising proteins under oil. *Journal of Crystal Growth* **168**, p175-180.

Dautant A, Meyer JB, Yariv J, Precigoux G, Sweet R, Kalb J, Frowlow F (1998). Structure of a monoclinic crystal form of cytochrome b1 (Bacterioferritin) from *E. coli*. *Acta. Cryst.* **D54**, p16-24

Ehrensberger A.H., Wilson D.K. (2004) Structural and catalytic diversityin the two family 11 aldo-keto reductases. *J. Mol. Biol.* **337**, p661-673

Eisenthal R., Peterson M.E., Daniel R.M., Danson M.J. (2006) The thermal behaviour of enzyme activity: implications for biotechnology. *Trends Biotechnol.* **24**, p289-292

Fleming T., Littlechild J. (1997) Sequence and structural comparison of thermophilic phosphoglycerate kinases with a mesophilic equivalent. *Comp. Biochem. Physiol. A* **118**, p439-451

Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta. Cryst*. **D60**, p2126-2132

Eriksson T, Bjorkman S, Roth B, Fyge A, Hoglund P (1995). Stereospecific determination, chiral inversion in vitro and pharmacokinetics in humans of the enantiomers of thalidomide. *Chirality* **7**, p44-52

French G.S., Wilson K.S. (1978) *Acta. Cryst.* **A34**, p517

Frolow F., Kalb A., Yariv J. (1994) Structure of a unique twofold symmetric haem-binding site. *Nat. Struct. Biol.* **1**, p453-460

Garman EF, Grime GW (2005) Elemental anaylsis of proteins by microPIXE. *Progress in Biophysics and Molecular Biology* **89**, p173-205.

Gasteiger E., Gattiker A., Hoogland C., Ivanyi I., Appel R.D., Bairoch A. (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res*. **31**, p3784-3788.

Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A. (2005) Protein Identification and Analysis Tools on the ExPASy Server; (In) John M. Walker (ed): *The Proteomics Protocols Handbook*, Humana Press. p571-607

Goldstein B., Blaxter M. (2002) Tardigrades. *Current Biology* **12**, p475

Grime, G. W. (2004) (In) M. Budnar (ed.): *Tenth International Conference on Particle Induced X-ray Emission*. Portoroz, Slovenia

Grimshaw C.E., Bohren K.M., Lai C.J., Gabbay K.H. (1995) Human aldose reductase: pK of tyrosine 48 reveals the preferred ionisation state for catalysis and inhibition. *Biochemistry* **34**, p14374-14384

Guy J.E. (2004) Ph.D. Thesis. University of Exeter, Exeter, U.K.

Guy J.E., Isupov M.N., Littlehild J.A. (2003a) Crystallisation and preliminary X-ray diffraction studies of a novel alcohol dehydrogenase from the hyperthermophilic archaeon *Aeropyrum pernix*. *Acta. Cryst.* **D59**, p174-176

Guy J.E., Isupov M.N., Littlechild J.A. (2003b) The structure of an alcohol dehydrogenase from the hyperthermophilic archaeon *Aeropyrum pernix*. *J. Mol. Biol.* **331**, p1041-1051

Hahn-Hagerdahl B., Linden T., Senac T., Skoog K. (1991) Ethanoic fermentation of pentoses in lignocellulose hydrolysates. *Appl. Biochem. Biotechnol.* **28**, p131-144

Hall T.A. (1999) A user friendly biological sequence alignment editor and analysis program for windows 95/98/NT. *Nucl. Acids. Symp. Ser.* **41**, p95-98

Hendrickson W.A. (1985) Analysis of Protein Structure from Diffraction Measurement at Multiple Wavelengths. *Trans. ACA* **21**.

Hensel R., Jakob I., Scheer H., Lottspeich F. (1992) Proteins from hyperthermophilic archaea: stability towards covalent modifications of the peptide chain. *Biochem. Soc. Symp.* **58**, p127-133

Hess D., Kruger K., Knappik A., Palm P., Hensel R. (1995) Dimeric 3-phosphoglycerate kinases from hyperthermophilic Archaea. Cloning, sequencing and expression of the 3-phosphoglycerate kinase gene of *Pyrococcus woesei* in *Escherichia coli* and characterization of the protein. Structural and functional comparison with the 3-phosphoglycerate kinase of *Methanothermus fervidus*. *Eur. J. Biochem.* **233**, p227-237

Hildebrandt P., Musidlowska A., Bornsheuer U.T., Altenbuchner J. (2002) Cloning, functional expression and biochemical characterization of a stereoselective alcohol dehydrogenase from *Pseudomonas fluorescens* DSM50106. *Appl. Microbiol. Biotechnol.* **59**, p483-487

Holland H.L (2000). Biotranformations; (In) *Encyclopedia of Microbiology*, Volume 1, Academic Press. p636-645

Huber R., Langworthy T.A., Konig H., Thomm M., Woese C.R., Sleytr U.B., Stetter K.O. (1986) *Thermotoga maritima* sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing upto 90°C. *Arch. Microbiol.* **144**, p324-333

Huismann G.W., Gray D (2002). Towards novel processes for the fine-chemical and pharmaceutical industries. *Curr. Opin. Biotechnol.* **13**, p352-358

Hummel W (1999). Large scale application of NAD(P)-dependant oxidoredutases: recent developments. *Trends Biotechnol.* **17**, p487-492

Ikeda H., Nonomiya T., Usami M., Ohta T., Omura S. (1999) Organisation of the biosynthetic cluster for the polyketide anthelmintic macrlide avermectin in *Streptomyces avermitilis*. *Proc. Nat. Acad. Sci.* **96**, p9509-9514

IuchiY., Kaneko T., Matsuki S., Ishii T., Ikeda Y., Uchida K., Fujii J. (2004) Carbonyl stress and detoxification ability in the male genital tract and testis of rats. *Histochem. Cell Biol.* **121**, p123-130

Jaenicke R. (1991) Protein stability and molecular adaptation to extreme conditions. *Eur. J. Biochem.* **202**, p715-728

Jez J.M., Bennet M.J., Schlegel B.P., Lewis M., Penning T.M. (1997) Comparative anatomy of the aldo-keto reductase superfamily. *Biochem. J.* **326**, p625-636

Ji A. and Gao P. (2001) Substrate specificity of Gluconobacter oxydans for production of 2,5-diketo-D-gluconic acid and 2-keto-L-gulonic acid in a multi enzyme system. *Appl. Biochem. Biotechnol.* **94**, p213-223

Jonsson K.I. (2007) Tardigrades as a potential model organism in space research. *Astrobiology* **7**, p757-766

Kabashima T., Li Y., Kanada N., Ito K., Yoshimoto T. (2001) Enchancement of the thermal stability of pyroglutamyl peptidase I by introduction of an intersubunit disulfide bond. *Biochim. Biophys. Acta.* **1547**, p214-220

Kabsch W (1976) *Acta. Cryst.* **A32**, p922-923

Kadir F.H.A., Al-Massad F.K., Moore G.R. (1992) Haem binding to horse spleen ferritin and its effect on the rate of iron release. *Biochem. J.* **282**, p867-870

Kador P.F. (1988) The role of aldose reductase in the development of diabetic complications. *Med. Res. Rev.* **8**, p325-352

Karle J. (1980). Some Developments in Anomalous Dispersion for the Structural Investigation of Macromolecular Systems in Biology. *International Journal of Quantum Chemistry: Quantum Biology Symposium* **7**, p357-367.

Kashefi K., Lovley D.R. (2003) Extending the upper temperature limit for life. *Science* **301**, p934

Keech A.M., Le Brun N.E, Wilson M.T., Andrews S.C., Moore G.R., Thomas A.J. (1997) Spectroscopic studies of cobalt(II) binding to *Escherichia coli* bacterioferritin. *J. Biol. Chem.* **272**, p422–429

Kendrew J.C., Bodo G., Dintzis H.M., Parrish R.G., Wyckoff H., Phillips D.C. (1958) A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* **181**, p662-666

Kikuchi A., Asai K. (1984) Reverse gyrase – a topoisomerase which introduces positive superhelical turns into DNA. *Nature* **309**, p677-681

Kirino H., Aoki M., Aoshima M., Hayashi Y., Ohba M., Yamagishi A., Wakagi T., Oshima T. (1994) Hydrophobic interaction at the subunit interface contributes to the thermostability of 3-isopropylmalate dehydrogenase from an extreme thermophile, *Thermus thermophilus*. *Eur. J. Biochem.* **220**, p275-281

Kirk O., Borchert T.V., Fuglsang C.C. (2002). Industrial Enzyme Applications. *Current Opinion in Biotechnology* **13**, p345-351

Knapp S., Kardinahl S., Hellgren N., Tibbelin G., Schafer G., Ladenstein R. (1999) Refined crystal structureof superoxide dismutase from the hyperthermophilic archaeon *Sulfolobus acidocaldarius* at 2.2Å resolution. *J. Mol. Biol.* **285**, p689-702

Laskowski R.A., MacArthur M.W., Moss D.S., Thornton J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26**, p283-291

Le Brun N.E., Wilson M.T., Andrews S.C., Guest J.R., Harrison P.M., Thomson A.J., Moore J.R. (1993). Kinetic and structural characterisation of an intermediate in the biomineralization of bacterioferritin. *FEBS Lett.* **333** p197-202

Le Brun N.E., Andrews S.C., Guest J.R., Harrison P.M., Moore G.R., Thomson A.J. (1995) Identification of the ferroxidase centre of *Escherichia coli* bacterioferritin. *Biochem. J.* **312**, p385-392

Lewin A., Moore G.R., Le Brun N.E. (2005) Formation of protein coated iron minerals. *Dalton Trans.* **22**, p3597-3610

Li W.T., Grayling R.A., Sandman K., Edmondson S., Shriver J.W., Reeve J.N. (1998) Thermodynamic stability of archaeal histones. *Biochemsitry* **37**, p10563-10572

Liu H.L., Zhou H.N., Xing W.M., Zhao J.F., Li S.X., Huang J.F., Bi R.C. (2004) 2.6Å resolution crystal structure of the bacterioferritin from Azotobacter vinelandii. *FEBS Letters* **573**, p93-98

Lovell S.C., Davis I.W., Arendall III W.B., de Bakker P.I.W., Word J.M., Prisant M.G., Richardson J.S., Richardson D.C. (2002) Structure validation by Cα geometry: φ/ψ and Cβ deviation. *Proteins: Structure, Function & Genetics* **50**, p437-450

Madigan M.T., Oren A. (1999) Thermophilic and halophilic extremophiles. *Curr. Opin. Microbiol.* **2**, p265-269

Makhatadze G.I., Privalov P.L. (1995) Energetics of protein stabilisation. *Adv. Protein Chem.* **47**, p307-425

Matsumura M., Becktel W.J., Matthews B.W. (1988) Hydrophobic stabilisation in T4 lysozyme determined directly by multiple substitutions of Ile3. *Nature* **334**, p406-410

Mayr P., Nidetzky B. (2002) Catalytic reaction profile for NADH-dependant reduction of aromatic aldehydes by xylose reductase from *Candida tenuis*. *Biochem. J.* **366**, p889-899

Michels P.C., Clark D.S. (1997) Pressure-enhanced activity and stability of a hyperthermophilic protease from a deepsea methanogen. *Appl. Environ. Microbiol.* **63**, p3985-3991

Mrabet N.T., Van den Broeck A., Van den brande I., Stanssens P., Laroche Y., Lambeir A.M., Matthijssens G., Jenkins J., Chiadmi M., van Tilbeurgh H., Rey F., Janin J., Quax W.J., Lasters I., De Maeyer M., Wodak S.J. (1992) Arginine residues as stabilising elements in proteins. *Biochemistry* **31**, p2239-2253

Murshudov G, Vagin A, Dodson E (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta. Cryst.* **D53**, p240-255

Nelson K.E., Clayton R.A., Gill S.R., Gwinn M.L., Dodson R.J., Haft D.H., Hickey E.K., Peterson J.D., Nelson W.C., Ketchum K.A., McDonald L., Utterback T.R., Malek J.A., Linher K.D., Garrett M.M., Stewart A.M., Cotton M.D., Pratt M.S., Phillips C.A., Richardson D., Heidelberg J., Sutton G.G., Fleischmann R.D., Eisen J.A., White O., Salzberg S.L., Smith H.O., Venter J.C., Fraser C.M. (1999) Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**, p323-329

Nidetzky B., Klimacek M., Mayr P. (2001) Transient-state and steady-state kinetic studies of the mechanism of NADH-dependant aldehyde reduction catalysed by xylose reductase from the yeast *Candida tenuis*. *Biochemistry* **40**, p10371-10381

Odagaki Y., Hayashi A., Okada K., Hirotsu K., Kabashima T., Ito K., Yoshimoto T., Tsuru D., Sato M., Clardy J. (1999) The crystal structure of pyroglutamyl peptidase I from *Bacillus amyloliquefaciens* reveals a new structure for a cysteine protease. *Structure 7*, p399-411

Otwinowski Z., Minor W. (1997) Processing of X-ray Diffraction Data Collected in Oscillation Mode. *Methods Enzymol.* **267**, p21839-21843

Pace C.N. (1992) Contribution of the hydrophobic effect to globular protein stability. *J. Mol. Biol.* **226**, p29-25

Pace N.C., Tanford C. (1968) Thermodynamics of the unfolding of β-lactglobulin A in aqueous urea solutions between 5°C and 55°C. *Biochemistry* **7**, p198-208

Paul D., Pandey G., Pandey J., Jain R.K. (2005) Accessing microbial diversity for bioremediation and environmental restoration. *Trends Biotechnol.* **23**, p135-142

Penning T.M. (1999) Molecular determinants of steroid recognition and catalysis in aldo-keto reductases: lessons from 3alpha hydroxysteroid dehydrogenase. *J. Steroid Biochem. Mol. Biol.* **69**, p211-225

Perrakis A., Sixma T. K., Wilson K.S., and Lamzin V. S. (1997) wARP: improvement and extension of crystallographic phases by weighted averaging of multiple refined dummy atomic models. *Acta. Cryst.* **D53**, p448-455,

Phipps B.M., Hoffmann A., StetterK.O., Baumeister W. (1991) A novel ATPase complex selectively accumulated upon heat shock is a major cellular component of thermophilic archaebacteria. *EMBO J.* **10**, p1711-1722

Reetz M.T. (2002) Directed evolution of selective enzymes and hybrid catalysts. *Tetrahedron* **58**. p6595–6602

Reetz M.T. (2004) Controlling the enantioselectivity of enzymes by directed evolution: Practical and theoretical ramifications. *PNAS.* **101**, p5716-5722

Reetz M.T., Bocola M., Carballeira J.D., Zha D., Vogel A. (2005) Expanding the range of substrate acceptance of enzymes: combinatorial active-site saturation test. *Angew. Chem. Int. Ed.* **44**, p4192-4196

Reetz M.T., Wang L.W., Bocola M. (2006) Directed Evolution of Enantioselective Enzymes: Iterative Cycles of CASTing for Probing Protein-Sequence Space. *Angew. Chem. Int. Ed.* **45**, p1236 –1241

Reetz M.T., Carballeira J.D. (2007) Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. *Nature protocols* **2**, p891-903

Reetz M.T., Puls M., Carballeira J.D., Vogel A., Jaeger K.E., Eggert T., Theil W., Bocola M., Otte N. (2007) Learning from directed evolution: Further lessons from theoretical investigations into cooperative mutations in lipase enantioselectivity. *ChemBioChem.* **8**, p106-112

Robb F.T., Clark D.S. (1999) Adaptation of proteins from hyperthermophiles to high pressure and high temperature. *J. Mol. Microbiol. Biotechnol.* **1**, p101-105

Rodriguez R., Chinea G., Lopez N., Pons T., Vriend G. (1998) Homology modeling, model and software evaluation: three related resources. *CABIOS* **14**, p523-528.

Rouchi M.A. (2002) Chiral roundup. *Chem. Eng. News* **80**, p43-50

Russel R.J., Ferguson J.M., Danson M.J., Hough D.W., Taylor G.I. (1997) The crystal structure of citrate synthase from the thermophilic archaeon *Pyrococcus furiosus* at 1.9Å resolution. *Biochemistry* **36**, p9983-9994

Sandman K., Krzycki J.A., Dobrinski B., Lurz R., Reeve J.N. (1990) HMf, a DNA-binding protein isolated from the hyperthermophilic archaeon *Methanothermus fervidus*, is most closely related to histones. *Proc. Nat. Acad. Sci.* **87**, p5788-5791

Sanli G., Blaber M. (2001) Structural assembly of the active site in an Aldo-Keto Reductase by NADPH cofactor. *J. Mol. Biol.* **309**, p1209-1218

Sanli G., Banta S., Anderson S., Blaber M. (2004) Structural alteration of cofactor specificity in corynebacterium 2,5-diketo-D-gluconic acid reductase. *Protein Sci.* **13**, p504-512

Schlegel B.P., Jez J.M., Penning T.M. (1998) Mutagenesis of 3 alpha-hydroysteroid dehydrogenase reveals a "push-pull" mechanism for proton transfer in aldo-keto reductases. *Biochemistry* **37**, p3538-3548

Seki K., Toyoshima M. (1998) Preserving tardigrades under pressure. *Nature* **395**, p853-854

Sheldrick GM (2008) A short history of SHELX. *Acta. Cryst.* **A64**, p112-122

Sherman F., Stewart J.W., Tsunasawa S. (1985) Methionine or not methionine at the beginning of a protein. *Bioessays* **3**, p27-31

Shirley B.A., Stanssens P., Hahn U., Pace C.N. (1992) Contribution of hydrogen bonding to the conformational stability of ribonuclease T1. *Biochemistry* **31**, p725-732

Singleton M.R., Isupov M.N., Littlechild J.A. (1999) Crystallisation and preliminary X-ray diffraction studies of pyrrolidone carboxyl peptidase from the hyperthermophilic archaeon *Thermococcus litoralis*. *Acta. Cryst.* **D55**, p703-703

Snyder L.R. (1974) Classification of the solvent properties of common liquids. *J. Chromatography* **A92**, p223-230

Sotriffer C.A., Kramer O., Klebe G. (2004) Probing flexibility and "induced-fit" phenomena in aldose reductase by comparative crystal structure analysis ad molecular dynamics simulations. *Proteins: structure, function and bioinformatics* **56**, p52-66

Stetter K.O. (1999) Extremophiles and their adaptations to hot environments. *FEBS letters* **452**, p22-25

Stillman T.J., Hempstead P.D., Artymiuk P.J., Andrews S.C., Hudson A.J., Treffry A., Guest J.R., Harrison P.M. (2001) The high resolution X-ray crystallographic of the ferritin (EcFtnA) of *Escherichia coli*; comparison with human H ferritin (HuHF) and the structures of the Fe^{3+} and Zn^{2+} derivatives. *J. Mol. Biol.* **307** p587-603

Straathof A.J.J., Panke S., Schmid A. (2002) The production of fine chemicals by biotransformations. *Curr. Opin. Biotechnol.* **13**, p548-556

Swartz L., Kuchinskas M., Li H., Poulos T.L., Lanzilotta W.N. (2006) Redox-dependent structural changes in the *Azotobacter vinelandii* bacterioferritin: new insights into the ferroxidase and iron transport mechanism. *Biochemistry* **45**, p4421-4428

Tarle I., Borhani D.W., Wilson D.K., Quiocho F.A., Petrash J.M. (1993) Probing the active site of human aldose reductase. Site directed mutagenesis of Asp-43, Tyr-48, Lys-77 and His-110. *J. Biol. Chem.* **268**, p25687-25693

Takikawa H., Ookhtens M., Stolz A., Kaplowitz N. (1987) Cyclical oxidation-reduction of the C3 position on bile acids catalyzed by 2 alpha-hydroxysteroiddehydrogenase. *J. Clin. Invest.* **80**, p861-866

Uchida K. (2000) Role of reactive aldehydes in cardiovascular disease. *Free Radical Biology and Medicine* **28**, p1685-1696

Vagin A, Teplyakov A. (1997) MOLREP: an automated program for molecular replacement. *J. Appl. Cryst.* **30**, p1022-1025

Vaguine A.A., Richelle J., Wodak S.J. (1999) SFCHECK: a unified set of procedure for evaluating the quality of macromolecular structure-factor data and their agreement with atomic model. *Acta. Cryst.* **D55**, p191-205

Van Eerde A., Walterink-van Loo S., Van der Oost J., Dijkstra B.W. (2006) Fortuitous structure determination of 'as isolated' *Escherichia coli* bacterioferritin in a novel crystal form. *Acta. Cryst.* **F62** p1061-1066

Vieille C., Hess J.M., Kelly R.M., Zeikus J.G. (1995) *xylA* cloning and sequencing and biochemical characterisation of xylose isomerase from *Themotoga neapolitana*. *Appl. Environ. Microbiol.* **61**, p1867-1875

Vieille C., Zeikus G.J. (2001) Hyperthermophilic enzymes: sources, uses, and molecular mechanisms of thermostability. *Microbiol. Mol. Biol. Rev.* **65**, p1-43

Villaroya A., Juan E., Egestad B., Jörnvall H. (1989). The primary structure of alcohol dehydrogenase from *Drosophila lebanonensis*. Extensive variation within insect 'short-chain'alcohol dehydrogenase lacking zinc. *Eur. J. Biochem.* **180**, p191-197

Williamson V.M., Paquain C.E. (1987). Homology of *Saccharomyces cerevisiae* ADH4 to an iron-activated alcohol dehydrogenase from *Zymomonas mobilis*. *Mol. Gen. Genet.* **209**, p374-381

Wierenga R.K. (2001) The TIM barrel fold: a versatile framework for efficient enzymes. *FEBS lett.* **492**, p193-198

Wong T.S., Roccatano D., Zacharias M., Schwaneberg U. (2006) A Statistical Analysis of Random Mutagenesis Methods Used for Directed Protein Evolution. *J. Mol. Biol.* **355**, p858–871

Wright H.T. (1991) Nonenzymatic deamidation of asparaginyl and glutamyl residues in proteins. *Crit. Rev. Biochem. Mol. Biol.* **26**, p1-52

Yamagata Y., Ogasahara K., Hioki Y., Lee S.J., Nakagawa A., Nakamura H., Ishida M., Kuramitsu S., Yutani K. (2001) Entropic stabilisation of the tryptophan synthase α-subunit from a hyperthermophilie, *Pyrococcus furiosus*. *J. Biol. Chem.* **276**, p11062-11071

Zaks A (2001). Industrial Biocatalysis. *Curr. Opin. Chem. Biol.* **5**, p130-136

Appendix 1. ExPASy output for TmAKR

Number of amino acids: 333

Molecular weight: 37538.2

Theoretical pI: 5.56

Amino acid composition:
Ala (A) 22   6.6%
Arg (R) 21   6.3%
Asn (N) 10   3.0%
Asp (D) 10   3.0%
Cys (C) 4   1.2%
Gln (Q) 7   2.1%
Glu (E) 43   2.9%
Gly (G) 29   8.7%
His (H) 5   1.5%
Ile (I) 23   6.9%
Leu (L) 33   9.9%
Lys (K) 25   7.5%
Met (M) 7   2.1%
Phe (F) 12   3.6%
Pro (P) 20   6.0%
Ser (S) 14   4.2%
Thr (T) 13   3.9%
Trp (W) 5   1.5%
Tyr (Y) 9   2.7%
Val (V) 21   6.3%

Total number of negatively charged residues (Asp + Glu): 53
Total number of positively charged residues (Arg + Lys): 46

Extinction coefficients:
Extinction coefficients are in units of M^{-1} cm^{-1}, at 280 nm measured in water.

Ext. coefficient   41160
Abs 0.1% (=1 g/l)   1.096, assuming ALL Cys residues appear as half cystines

Ext. coefficient   40910
Abs 0.1% (=1 g/l)   1.090, assuming NO Cys residues appear as half cystines

Appendix 2. Gel Filtration Calibration Curve
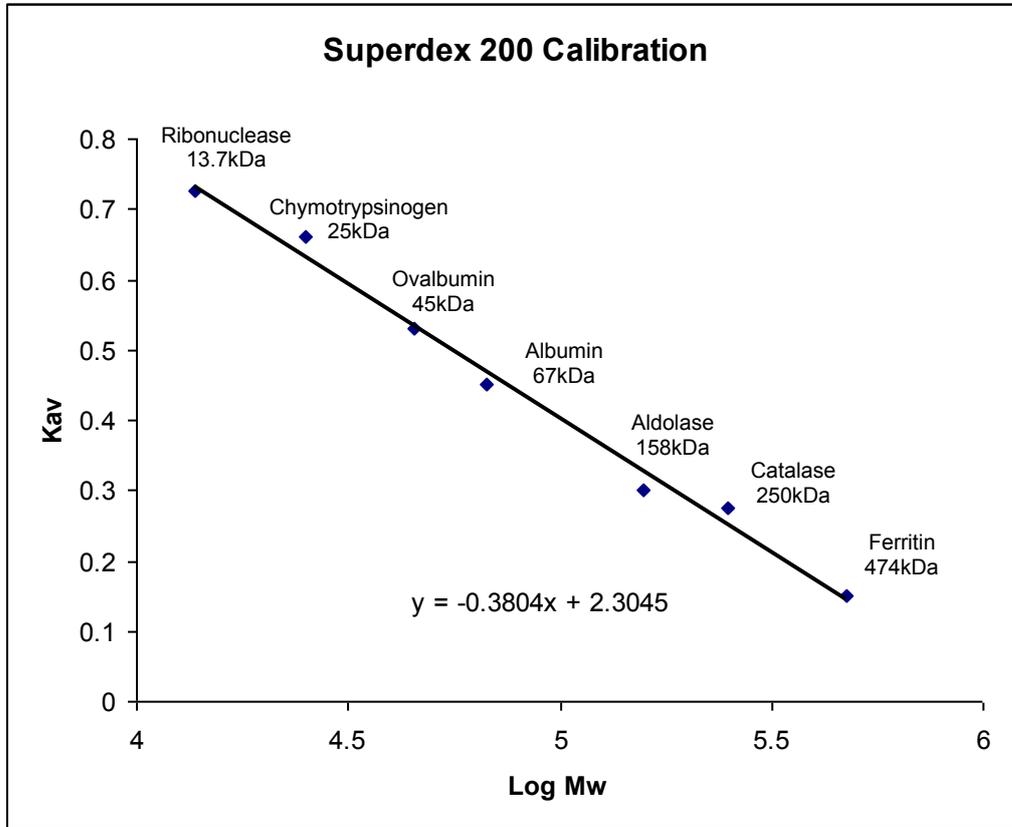
Calibration curve for Superdex200 gel filtration column.

Appendix 3. Substrate structures

1,2,3,6-tetrahydrobenzaldehyde
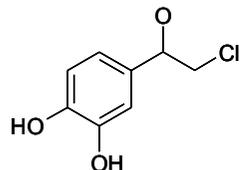
2-chloro-3',4'-dihydroxyacetophenone
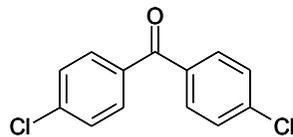
dichlorobenzophenone
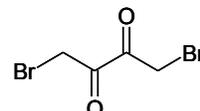
1,4-dibromo-2,3-butanedione
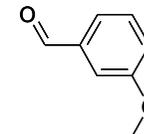
m-anisaldehyde
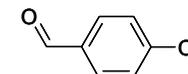
p-anisaldehyde
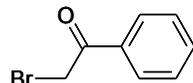
2-bromoacetophenone
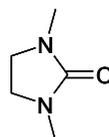
1,3-dimethylimidazolidinone
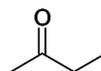
2-butanone
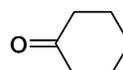
cyclohexanone
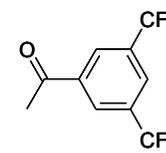
3,5-(bistrifluromethyl)acetophenone
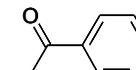
acetophenone
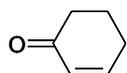
2-cyclohexen-1-one
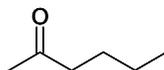
2-hexanone
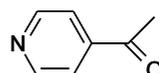
4-acetylpyridine
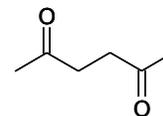
acetonyl acetone
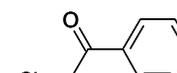
2-chloroacetophenone
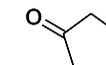
cyclopentanone
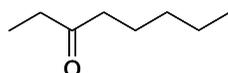
3-octanone
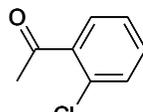
2'-chloroacetophenone
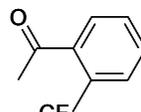
2'-(trifluromethyl)acetophenone
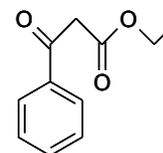
ethyl benzoylacetate
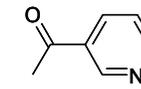
3-acetylpyridine
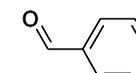
benzaldehyde
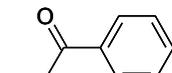
propiophenone
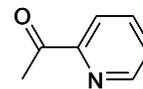
2-acetyl pyridine
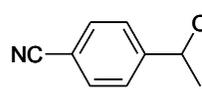
4-acetylbenzonitrile
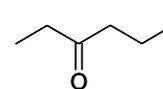
3-hexanone

acetone

Appendix 4. Crystallisation screen contents

MDL Structure Screen 1

|	Additive	Buffer	Precipitant
1	0.02M Calcium chloride dihydrate	0.1M Na Acetate trihydrate pH 4.6	30% v/v 2-methyl-2,4 pentanediol
2	0.2M Ammonium acetate	0.1M Na Acetate trihydrate pH 4.6	30% w/v PEG 4000
3	0.2M Ammonium sulfate	0.1M Na acetate trihydrate pH 4.6	25% w/v PEG 4000
4	None	0.1M Na acetate trihydrate pH 4.6	2.0M Sodium formate
5	None	0.1M Na acetate trihydrate pH 4.6	2.0M Ammonium sulfate
6	None	0.1M Na acetate trihydrate pH 4.6	8% w/v PEG 4000
7	0.2M Ammonium acetate	0.1M tri-sodium citrate dihydrate pH 5.6	30% w/v PEG 4000
8	0.2M Ammonium acetate	0.1M tri-sodium citrate dihydrate pH 5.6	30% v/v 2-methyl-2,4-pentanediol
9	None	0.1M tri-Sodium citrate dihydrate pH 5.6	20% w/v 2-propanol, 20% w/v PEG 4000
10	None	0.1M Na Citrate pH 5.6	1.0M Ammonium dihydrogen phosphate
11	0.2M Calcium chloride dihydrate	0.1M Na acetate trihydrate pH 4.6	20% v/v 2-propanol
12	None	0.1M Na Cacodylate pH 6.5	1.4M Na acetate trihydrate
13	0.2M tri-sodium citrate dihydrate	0.1M Na Cacodylate pH 6.5	30% v/v 2-propanol
14	0.2M Ammoniium sulfate	0.1M Na Cacodylate pH 6.5	30% w/v PEG 8000
15	0.2M Magnesium acetate tetrahydrate	0.1M Na Cacodylate pH 6.5	20% PEG 8000
16	0.2M Magnesium acetate tetrahydrate	0.1M Na Cacodylate pH 6.5	30% v/v 2-methyl-2,4-pentanediol
17	None	0.1M Imidazole pH 6.5	1.0M Sodium acetate trihydrate
18	0.2M Sodium acetate trihydrate	0.1M Na Cacodylate pH 6.5	30% w/v PEG 8000
19	0.2M Zinc acetate dihydrate	0.1M Na Cacodylate pH 6.5	18% w/v PEG 8000
20	0.2M Calcium acetate hydrate	0.1M Na Cacodylate pH 6.5	18% w/v PEG 8000
21	0.2M tri-sodium citrate dihydrate	0.1M Na Hepes pH 7.5	30% v/v 2-methyl-2,4-pentanediol
22	0.2M Magnesium chloride hexahydrate	0.1M Na Hepes pH 7.5	30% v/v 2-propanol
23	0.2M Calcium chloride dihydrate	0.1M Na Hepes pH 7.5	28% v/v PEG 400
24	0.2M Magnesium chloride hexahydrate	0.1M Na Hepes pH 7.5	30% v/v PEG 400
25	0.2M tri-sodium citrate dihydrate	0.1M Na Hepes pH 7.5	20% v/v 2-propanol
26	None	0.1M Na Hepes pH 7.5	0.8M K, Na tartrate tetrahydrate
27	None	0.1M Na Hepes pH 7.5	1.5M Lithium sulfate monohydrate
28	None	0.1M Na Hepes pH 7.5	0.8M Na dihydrogen phosphate
29	None	0.1M Na Hepes pH 7.5	1.4M tri-Sodium citrate dihydrate
30	None	0.1M Na Hepes pH 7.5	2% v/v PEG 400, 2.0M Amm sulfate
31	None	0.1M Na Hepes pH 7.5	10% v/v 2-propanol, 20% w/v PEG 4000
32	None	0.1M Tris-HCl pH 8.5	2.0M Ammoniium sulfate
33	0.2M Magnesium chloride hexahydrate	0.1M Tris-HCl pH 8.5	30% w/v PEG 4000
34	0.2M tri-sodum citrate dihydrate	0.1M Tris-HCl pH 8.5	30% v/v PEG 400
35	0.2M Lithium sulfate monohydrate	0.1M Tris-HCl pH 8.5	30% w/v PEG 4000
36	0.2M Ammonium acetate	0.1M Tris-HCl pH 8.5	30% v/v 2-propanol
37	0.2M Sodium acetate trihydrate	0.1M Tris-HCl pH 8.5	30% w/v PEG 4000
38	None	0.1M Tris-HCl pH 8.5	8% w/v PEG 8000
39	None	0.1M Tris-HCl pH 8.5	2.0M Ammonium dihydrogen phosphate
40	None	None	0.4M K, Na Tartrate tetrahydrate

41	None	None	0.4M Ammonium dihydrogen phosphate
42	0.2M Ammonium sulfate	None	30% w/v PEG 8000
43	0.2M Ammonium sulfate	None	30% w/v PEG 4000
44	None	None	2.0M Ammonium sulfate
45	None	None	4.0M Sodium formate
46	0.05M Potassium dihydrogen phosphate	None	20% w/v PEG 8000
47	None	None	30% w/v PEG 1500
48	None	None	0.2M Magnesium formate
49	1.0M Lithium sulfate monohydrate	None	2% w/v PEG 8000
50	0.5M Lithium sulfate monohydrate	None	15% w/v PEG 8000

Components of crystallisation screen solutions in Structure Screen 1 (Molecular Dimensions Ltd)

MDL Structure Screen 2

	Additive	Buffer	Precipitant
1	0.1M Sodium chloride	0.1M Bicine pH 9.0	30% w/v PEG monomethylether 550
2	None	0.1M Bicine pH 9.0	2.0M Magnesium chloride hexahydrate
3	2% w/v Dioxane	0.1M Bicine pH 9.0	10% w/v PEG 20,000
4	0.2M Magnesium chloride hexahydrate	0.1M Tris pH 8.5	3.4M 1,6 Hexanediol
5	None	0.1M Tris pH 8.5	25% v/v tert-Butanol
6	0.01M Nickel chloride hexahydrate	0.1M Tris pH 8.5	1.0M Lithium sulfate
7	1.5M Ammonium sulfate	0.1M Tris pH 8.5	12% v/v Glycerol
8	0.2M Ammonium phosphate monobasic	0.1M Tris pH 8.5	50% v/v MPD
9	None	0.1M Tris pH 8.5	20% v/v Ethanol
10	0.01M Nickel chloride hexahydrate	0.1M Tris pH 8.5	20% w/v PEG monomethylether 2000
11	0.5M Ammonium sulfate	0.1M Hepes pH 7.5	30% v/v MPD
12	None	0.1M Hepes pH 7.5	10% w/v PEG 6000, 5% v/v MPD
13	None	0.1M Hepes pH 7.5	20% v/v Jeffamine M-600
14	0.1M Sodium chloride	0.1M Hepes pH 7.5	1.6M Ammonium sulfate
15	None	0.1M Hepes pH 7.5	2.0M Ammonium formate
16	0.05M Cadmium sulfate octahydrate	0.1M Hepes pH 7.5	1.0M Sodium acetate
17	None	0.1M Hepes pH 7.5	70% v/v MPD
18	None	0.1M Hepes pH 7.5	4.3M Sodium chloride
19	None	0.1M Hepes pH 7.5	10% w/v PEG 8000
20	None	0.1M MES pH 6.5	1.6M Magnesium sulfate heptahydrate
21	0.1M Na phosphate monobasic	0.1M MES pH 6.5	2.0M Sodium Chloride
22	None	0.1M MES pH 6.5	12% w/v PEG 20,000
23	1.6M Ammonium sulfate	0.1M MES pH 6.5	10% v/v Dioxane
24	0.05M Cesium chloride	0.1M MES pH 6.5	30% v/v Jeffamine M-600
25	0.01M Cobalt chloride hexahydrate	0.1M MES pH 6.5	1.8M Ammonium sulfate
26	0.2M Ammonium sulfate	0.1M MES pH 6.5	30% w/v PEG monomethylether 5000
27	0.01M Zinc sulfate heptahydrate	0.1M MES pH 6.5	25% v/v PEG monomethylether 550
28	None	0.1M Hepes pH 7.5	20% w/v PEG 10,000
29	0.2M K/Na Tartrate	0.1M Sodium citrate pH 5.6	2.0M Ammonium sulfate

30	0.5M Ammonium sulfate	0.1M Sodium citrate pH 5.6	1.0M Lithium sulfate
31	0.5M Sodium chloride	0.1M Sodium citrate pH 5.6	4% v/v polyethyleneimine
32	None	0.1M Sodium citrate pH 5.6	35% v/v tert-butanol
33	0.01M Ferric chloride hexahydrate	0.1M Sodium citrate pH 5.6	10% v/v Jeffamine M-600
34	0.01M Manganese chloride tetrahydrate	0.1M Sodium citrate pH 5.6	2.5M 1,6 Hexanediol
35	None	0.1M Sodium acetate pH 4.6	2.0M Sodium chloride
36	0.2M Sodium Chloride	0.1M Sodium acetate pH 4.6	30% v/v MPD
37	0.01M Cobalt Chloride hexahydrate	0.1M Sodium acetate pH 4.6	1.0M 1,6 Hexanediol
38	0.1M Cadmium chloride	0.1M Sodium acetate pH 4.6	30% v/v PEG 400
39	0.2M Ammonium sulfate	0.1M Sodium acetate pH 4.6	30% w/v PEG monomethylether 2000
40	2.0M Sodium Chloride	None	10% w/v PEG 6000
41	0.01M Cetyl trimethyl ammoniumbromide	None	0.5M Sodium chloride 0.1M Magnesium chloride hexahydrate
42	None	None	25% v/v Ethylene glycol
43	None	None	35% v/v Dioxane
44	2.0M Ammonium Sulfate	None	5% v/v Isopropanol
45	None	None	1.0M Imidazole pH 7.0
46	None	None	10% w/v PEG 1000, 10% w/v PEG 8000
47	1.5M Sodium Chloride	None	10% v/v Ethanol
48	None	None	1.6M Sodium citrate pH 6.5
49	15% w/v Polyvinylpyrolidone	None	None
50	2.0M Urea	None	None

Components of crystallisation screen solutions in Structure Screen 2 (Molecular Dimensions Ltd)

pHclear I

No.	Buffer	Precipitant	pH	No.	Buffer	Precipitant	pH
1	0.1 M Citric acid	1.0 M NaCl	4.0	49	0.1 M Citric acid	0.8 M NH_4SO_4	4.0
2	0.1 M Citric acid	1.0 M NaCl	5.0	50	0.1 M Citric acid	0.8 M NH_4SO_4	5.0
3	0.1 M MES	1.0 M NaCl	6.0	51	0.1 M MES	0.8 M NH_4SO_4	6.0
4	0.1 M HEPES	1.0 M NaCl	7.0	52	0.1 M HEPES	0.8 M NH_4SO_4	7.0
5	0.1 M TRIS	1.0 M NaCl	8.0	53	0.1 M TRIS	0.8 M NH_4SO_4	8.0
6	0.1 M BICINE	1.0 M NaCl	9.0	54	0.1 M BICINE	0.8 M NH_4SO_4	9.0
7	0.1 M Citric acid	2.0 M NaCl	4.0	55	0.1 M Citric acid	1.6 M NH_4SO_4	4.0
8	0.1 M Citric acid	2.0 M NaCl	5.0	56	0.1 M Citric acid	1.6 M NH_4SO_4	5.0
9	0.1 M MES	2.0 M NaCl	6.0	57	0.1 M MES	1.6 M NH_4SO_4	6.0
10	0.1 M HEPES	2.0 M NaCl	7.0	58	0.1 M HEPES	1.6 M NH_4SO_4	7.0
11	0.1 M TRIS	2.0 M NaCl	8.0	59	0.1 M TRIS	1.6 M NH_4SO_4	8.0
12	0.1 M BICINE	2.0 M NaCl	9.0	60	0.1 M BICINE	1.6 M NH_4SO_4	9.0
13	0.1 M Citric acid	3.0 M NaCl	4.0	61	0.1 M Citric acid	2.4 M NH_4SO_4	4.0
14	0.1 M Citric acid	3.0 M NaCl	5.0	62	0.1 M Citric acid	2.4 M NH_4SO_4	5.0
15	0.1 M MES	3.0 M NaCl	6.0	63	0.1 M MES	2.4 M NH_4SO_4	6.0
16	0.1 M HEPES	3.0 M NaCl	7.0	64	0.1 M HEPES	2.4 M NH_4SO_4	7.0
17	0.1 M TRIS	3.0 M NaCl	8.0	65	0.1 M TRIS	2.4 M NH_4SO_4	8.0
18	0.1 M BICINE	3.0 M NaCl	9.0	66	0.1 M BICINE	2.4 M NH_4SO_4	9.0
19	0.1 M Citric acid	4.0 M NaCl	4.0	67	0.1 M Citric acid	3.2 M NH_4SO_4	4.0
20	0.1 M Citric acid	4.0 M NaCl	5.0	68	0.1 M Citric acid	3.2 M NH_4SO_4	5.0
21	0.1 M MES	4.0 M NaCl	6.0	69	0.1 M MES	3.2 M NH_4SO_4	6.0
22	0.1 M HEPES	4.0 M NaCl	7.0	70	0.1 M HEPES	3.2 M NH_4SO_4	7.0
23	0.1 M TRIS	4.0 M NaCl	8.0	71	0.1 M TRIS	3.2 M NH_4SO_4	8.0
24	0.1 M BICINE	4.0 M NaCl	9.0	72	0.1 M BICINE	3.2 M NH_4SO_4	9.0
25	0.1 M Citric acid	5 %(w/v) PEG 6000	4.0	73	0.1 M Citric acid	10 %(v/v) MPD	4.0
26	0.1 M Citric acid	5 %(w/v) PEG 6000	5.0	74	0.1 M Sodium acetate	10 %(v/v) MPD	5.0
27	0.1 M MES	5 %(w/v) PEG 6000	6.0	75	0.1 M MES	10 %(v/v) MPD	6.0
28	0.1 M HEPES	5 %(w/v) PEG 6000	7.0	76	0.1 M HEPES	10 %(v/v) MPD	7.0
29	0.1 M TRIS	5 %(w/v) PEG 6000	8.0	77	0.1 M TRIS	10 %(v/v) MPD	8.0
30	0.1 M BICINE	5 %(w/v) PEG 6000	9.0	78	0.1 M BICINE	10 %(v/v) MPD	9.0
31	0.1 M Citric acid	10 %(w/v) PEG 6000	4.0	79	0.1 M Citric acid	20 %(v/v) MPD	4.0
32	0.1 M Citric acid	10 %(w/v) PEG 6000	5.0	80	0.1 M Sodium acetate	20 %(v/v) MPD	5.0
33	0.1 M MES	10 %(w/v) PEG 6000	6.0	81	0.1 M MES	20 %(v/v) MPD	6.0
34	0.1 M HEPES	10 %(w/v) PEG 6000	7.0	82	0.1 M HEPES	20 %(v/v) MPD	7.0
35	0.1 M TRIS	10 %(w/v) PEG 6000	8.0	83	0.1 M TRIS	20 %(v/v) MPD	8.0
36	0.1 M BICINE	10 %(w/v) PEG 6000	9.0	84	0.1 M BICINE	20 %(v/v) MPD	9.0
37	0.1 M Citric acid	20 %(w/v) PEG 6000	4.0	85	0.1 M Citric acid	40 %(v/v) MPD	4.0
38	0.1 M Citric acid	20 %(w/v) PEG 6000	5.0	86	0.1 M Sodium acetate	40 %(v/v) MPD	5.0
39	0.1 M MES	20 %(w/v) PEG 6000	6.0	87	0.1 M MES	40 %(v/v) MPD	6.0
40	0.1 M HEPES	20 %(w/v) PEG 6000	7.0	88	0.1 M HEPES	40 %(v/v) MPD	7.0
41	0.1 M TRIS	20 %(w/v) PEG 6000	8.0	89	0.1 M TRIS	40 %(v/v) MPD	8.0
42	0.1 M BICINE	20 %(w/v) PEG 6000	9.0	90	0.1 M BICINE	40 %(v/v) MPD	9.0
43	0.1 M Citric acid	30 %(w/v) PEG 6000	4.0	91	0.1 M Citric acid	65 %(v/v) MPD	4.0
44	0.1 M Citric acid	30 %(w/v) PEG 6000	5.0	92	0.1 M Sodium acetate	65 %(v/v) MPD	5.0
45	0.1 M MES	30 %(w/v) PEG 6000	6.0	93	0.1 M MES	65 %(v/v) MPD	6.0
46	0.1 M HEPES	30 %(w/v) PEG 6000	7.0	94	0.1 M HEPES	65 %(v/v) MPD	7.0
47	0.1 M TRIS	30 %(w/v) PEG 6000	8.0	95	0.1 M TRIS	65 %(v/v) MPD	8.0
48	0.1 M BICINE	30 %(w/v) PEG 6000	9.0	96	0.1 M BICINE	65 %(v/v) MPD	9.0

Components of crystallisation screen solutions in pHclear I (Qiagen)

pHclear II

No.	Salt	Buffer	Precipitant	pH	No.	Salt	Buffer	Precipitant	pH
1	1.0 M LiCl	0.1 M Citric acid		4.0	49			0.8 M Na/K phosphate	5.0
2	1.0 M LiCl	0.1 M Citric acid		5.0	50			0.8 M Na/K phosphate	5.6
3	1.0 M LiCl	0.1 M MES		6.0	51			0.8 M Na/K phosphate	6.3
4	1.0 M LiCl	0.1 M HEPES		7.0	52			0.8 M Na/K phosphate	6.9
5	1.0 M LiCl	0.1 M TRIS		8.0	53			0.8 M Na/K phosphate	7.5
6	1.0 M LiCl	0.1 M BICINE		9.0	54			0.8 M Na/K phosphate	8.2
7	1.0 M LiCl	0.1 M Citric acid	10 %(w/v) PEG 6000	4.0	55			1.0 M Na/K phosphate	5.0
8	1.0 M LiCl	0.1 M Citric acid	10 %(w/v) PEG 6000	5.0	56			1.0 M Na/K phosphate	5.6
9	1.0 M LiCl	0.1 M MES	10 %(w/v) PEG 6000	6.0	57			1.0 M Na/K phosphate	6.3
10	1.0 M LiCl	0.1 M HEPES	10 %(w/v) PEG 6000	7.0	58			1.0 M Na/K phosphate	6.9
11	1.0 M LiCl	0.1 M TRIS	10 %(w/v) PEG 6000	8.0	59			1.0 M Na/K phosphate	7.5
12	1.0 M LiCl	0.1 M BICINE	10 %(w/v) PEG 6000	9.0	60			1.0 M Na/K phosphate	8.2
13	1.0 M LiCl	0.1 M Citric acid	20 %(w/v) PEG 6000	4.0	61			1.4 M Na/K phosphate	5.0
14	1.0 M LiCl	0.1 M Citric acid	20 %(w/v) PEG 6000	5.0	62			1.4 M Na/K phosphate	5.6
15	1.0 M LiCl	0.1 M MES	20 %(w/v) PEG 6000	6.0	63			1.4 M Na/K phosphate	6.3
16	1.0 M LiCl	0.1 M HEPES	20 %(w/v) PEG 6000	7.0	64			1.4 M Na/K phosphate	6.9
17	1.0 M LiCl	0.1 M TRIS	20 %(w/v) PEG 6000	8.0	65			1.4 M Na/K phosphate	7.5
18	1.0 M LiCl	0.1 M BICINE	20 %(w/v) PEG 6000	9.0	66			1.4 M Na/K phosphate	8.2
19	1.0 M LiCl	0.1 M Citric acid	30 %(w/v) PEG 6000	4.0	67			1.8 M Na/K phosphate	5.0
20	1.0 M LiCl	0.1 M Citric acid	30 %(w/v) PEG 6000	5.0	68			1.8 M Na/K phosphate	5.6
21	1.0 M LiCl	0.1 M MES	30 %(w/v) PEG 6000	6.0	69			1.8 M Na/K phosphate	6.3
22	1.0 M LiCl	0.1 M HEPES	30 %(w/v) PEG 6000	7.0	70			1.8 M Na/K phosphate	6.9
23	1.0 M LiCl	0.1 M TRIS	30 %(w/v) PEG 6000	8.0	71			1.8 M Na/K phosphate	7.5
24	1.0 M LiCl	0.1 M BICINE	30 %(w/v) PEG 6000	9.0	72			1.8 M Na/K phosphate	8.2
25		0.1 M Citric acid	5 %(v/v) Isopropanol	4.0	73			1.0 M Sodium malonate	4.0
26		0.1 M Citric acid	5 %(v/v) Isopropanol	5.0	74			1.5 M Sodium malonate	4.0
27		0.1 M MES	5 %(v/v) Isopropanol	6.0	75			1.9 M Sodium malonate	4.0
28		0.1 M HEPES	5 %(v/v) Isopropanol	7.0	76			2.4 M Sodium malonate	4.0
29		0.1 M TRIS	5 %(v/v) Isopropanol	8.0	77			2.9 M Sodium malonate	4.0
30		0.1 M BICINE	5 %(v/v) Isopropanol	9.0	78			3.4 M Sodium malonate	4.0
31		0.1 M Citric acid	10 %(v/v) Isopropanol	4.0	79			1.0 M Sodium malonate	5.0
32		0.1 M Citric acid	10 %(v/v) Isopropanol	5.0	80			1.5 M Sodium malonate	5.0
33		0.1 M MES	10 %(v/v) Isopropanol	6.0	81			1.9 M Sodium malonate	5.0
34		0.1 M HEPES	10 %(v/v) Isopropanol	7.0	82			2.4 M Sodium malonate	5.0
35		0.1 M TRIS	10 %(v/v) Isopropanol	8.0	83			2.9 M Sodium malonate	5.0
36		0.1 M BICINE	10 %(v/v) Isopropanol	9.0	84			3.4 M Sodium malonate	5.0
37		0.1 M Citric acid	20 %(v/v) Isopropanol	4.0	85			1.0 M Sodium malonate	6.0
38		0.1 M Citric acid	20 %(v/v) Isopropanol	5.0	86			1.5 M Sodium malonate	6.0
39		0.1 M MES	20 %(v/v) Isopropanol	6.0	87			1.9 M Sodium malonate	6.0
40		0.1 M HEPES	20 %(v/v) Isopropanol	7.0	88			2.4 M Sodium malonate	6.0
41		0.1 M TRIS	20 %(v/v) Isopropanol	8.0	89			2.9 M Sodium malonate	6.0
42		0.1 M BICINE	20 %(v/v) Isopropanol	9.0	90			3.4 M Sodium malonate	6.0
43		0.1 M Citric acid	30 %(v/v) Isopropanol	4.0	91			1.0 M Sodium malonate	7.0
44		0.1 M Citric acid	30 %(v/v) Isopropanol	5.0	92			1.5 M Sodium malonate	7.0
45		0.1 M MES	30 %(v/v) Isopropanol	6.0	93			1.9 M Sodium malonate	7.0
46		0.1 M HEPES	30 %(v/v) Isopropanol	7.0	94			2.4 M Sodium malonate	7.0
47		0.1 M TRIS	30 %(v/v) Isopropanol	8.0	95			2.9 M Sodium malonate	7.0
48		0.1 M BICINE	30 %(v/v) Isopropanol	9.0	96			3.4 M Sodium malonate	7.0

Components of crystallisation screen solutions in pHclear II (Qiagen)

PEG Screen I

No.	Salt	Buffer	Precipitant
1		0.1 M Sodium acetate pH 4.6	40 %(v/v) PEG 200
2		0.1 M Sodium acetate pH 4.6	30 %(v/v) PEG 300
3		0.1 M Sodium acetate pH 4.6	30 %(v/v) PEG 400
4		0.1 M Sodium acetate pH 4.6	25 %(v/v) PEG 550 MME
5		0.1 M Sodium acetate pH 4.6	25 %(w/v) PEG 1000
6		0.1 M Sodium acetate pH 4.6	25 %(w/v) PEG 2000 MME
7		0.1 M MES pH 6.5	40 %(v/v) PEG 200
8		0.1 M MES pH 6.5	30 %(v/v) PEG 300
9		0.1 M MES pH 6.5	30 %(v/v) PEG 400
10		0.1 M MES pH 6.5	25 %(v/v) PEG 550 MME
11		0.1 M MES pH 6.5	25 %(w/v) PEG 1000
12		0.1 M MES pH 6.5	25 %(w/v) PEG 2000 MME
13		0.1 M HEPES pH 7.5	40 %(v/v) PEG 200
14		0.1 M HEPES pH 7.5	30 %(v/v) PEG 300
15		0.1 M HEPES pH 7.5	30 %(v/v) PEG 400
16		0.1 M HEPES pH 7.5	25 %(v/v) PEG 550 MME
17		0.1 M HEPES pH 7.5	25 %(w/v) PEG 1000
18		0.1 M HEPES pH 7.5	25 %(w/v) PEG 2000 MME
19		0.1 M TRIS.HCl pH 8.5	40 %(v/v) PEG 200
20		0.1 M TRIS.HCl pH 8.5	30 %(v/v) PEG 300
21		0.1 M TRIS.HCl pH 8.5	30 %(v/v) PEG 400
22		0.1 M TRIS.HCl pH 8.5	25 %(v/v) PEG 550 MME
23		0.1 M TRIS.HCl pH 8.5	25 %(w/v) PEG 1000
24		0.1 M TRIS.HCl pH 8.5	25 %(w/v) PEG 2000 MME
25		0.1 M Sodium acetate pH 4.6	25 %(w/v) PEG 3000
26		0.1 M Sodium acetate pH 4.6	25 %(w/v) PEG 4000
27		0.1 M Sodium acetate pH 4.6	25 %(w/v) PEG 6000
28		0.1 M Sodium acetate pH 4.6	25 %(w/v) PEG 8000
29		0.1 M Sodium acetate pH 4.6	20 %(w/v) PEG 10000
30		0.1 M Sodium acetate pH 4.6	15 %(w/v) PEG 20000
31		0.1 M MES pH 6.5	25 %(w/v) PEG 3000
32		0.1 M MES pH 6.5	25 %(w/v) PEG 4000
33		0.1 M MES pH 6.5	25 %(w/v) PEG 6000
34		0.1 M MES pH 6.5	25 %(w/v) PEG 8000
35		0.1 M MES pH 6.5	20 %(w/v) PEG 10000
36		0.1 M MES pH 6.5	15 %(w/v) PEG 20000
37		0.1 M HEPES pH 7.5	25 %(w/v) PEG 3000
38		0.1 M HEPES pH 7.5	25 %(w/v) PEG 4000
39		0.1 M HEPES pH 7.5	25 %(w/v) PEG 6000
40		0.1 M HEPES pH 7.5	25 %(w/v) PEG 8000
41		0.1 M HEPES pH 7.5	20 %(w/v) PEG 10000
42		0.1 M HEPES pH 7.5	15 %(w/v) PEG 20000
43		0.1 M TRIS.HCl pH 8.5	25 %(w/v) PEG 3000
44		0.1 M TRIS.HCl pH 8.5	25 %(w/v) PEG 4000
45		0.1 M TRIS.HCl pH 8.5	25 %(w/v) PEG 6000
46		0.1 M TRIS.HCl pH 8.5	25 %(w/v) PEG 8000
47		0.1 M TRIS.HCl pH 8.5	20 %(w/v) PEG 10000
48		0.1 M TRIS.HCl pH 8.5	15 %(w/v) PEG 20000

Cont...

PEGs Screen I cont...

No.	Salt	Buffer	Precipitant
49	0.2 M Sodium fluoride		20 %(w/v) PEG 3350
50	0.2 M Potassium fluoride		20 %(w/v) PEG 3350
51	0.2 M Ammonium fluoride		20 %(w/v) PEG 3350
52	0.2 M Lithium chloride		20 %(w/v) PEG 3350
53	0.2 M Magnesium chloride		20 %(w/v) PEG 3350
54	0.2 M Sodium chloride		20 %(w/v) PEG 3350
55	0.2 M Calcium chloride		20 %(w/v) PEG 3350
56	0.2 M Potassium chloride		20 %(w/v) PEG 3350
57	0.2 M Ammonium chloride		20 %(w/v) PEG 3350
58	0.2 M Sodium iodide		20 %(w/v) PEG 3350
59	0.2 M Potassium iodide		20 %(w/v) PEG 3350
60	0.2 M Ammonium iodide		20 %(w/v) PEG 3350
61	0.2 M Sodium thiocyanate		20 %(w/v) PEG 3350
62	0.2 M Potassium thiocyanate		20 %(w/v) PEG 3350
63	0.2 M Lithium nitrate		20 %(w/v) PEG 3350
64	0.2 M Magnesium nitrate		20 %(w/v) PEG 3350
65	0.2 M Sodium nitrate		20 %(w/v) PEG 3350
66	0.2 M Potassium nitrate		20 %(w/v) PEG 3350
67	0.2 M Ammonium nitrate		20 %(w/v) PEG 3350
68	0.2 M Magnesium formate		20 %(w/v) PEG 3350
69	0.2 M Sodium formate		20 %(w/v) PEG 3350
70	0.2 M Potassium formate		20 %(w/v) PEG 3350
71	0.2 M Ammonium formate		20 %(w/v) PEG 3350
72	0.2 M Lithium acetate		20 %(w/v) PEG 3350
73	0.2 M Magnesium acetate		20 %(w/v) PEG 3350
74	0.2 M Zinc acetate		20 %(w/v) PEG 3350
75	0.2 M Sodium acetate		20 %(w/v) PEG 3350
76	0.2 M Calcium acetate		20 %(w/v) PEG 3350
77	0.2 M Potassium acetate		20 %(w/v) PEG 3350
78	0.2 M Ammonium acetate		20 %(w/v) PEG 3350
79	0.2 M Lithium sulfate		20 %(w/v) PEG 3350
80	0.2 M Magnesium sulfate		20 %(w/v) PEG 3350
81	0.2 M Sodium sulfate		20 %(w/v) PEG 3350
82	0.2 M Potassium sulfate		20 %(w/v) PEG 3350
83	0.2 M Ammonium sulfate		20 %(w/v) PEG 3350
84	0.2 M di-Sodium tartate		20 %(w/v) PEG 3350
85	0.2 M K/Na tartrate		20 %(w/v) PEG 3350
86	0.2 M di-Ammonium tartrate		20 %(w/v) PEG 3350
87	0.2 M Sodium phosphate		20 %(w/v) PEG 3350
88	0.2 M di-Sodium phosphate		20 %(w/v) PEG 3350
89	0.2 M Potassium phosphate		20 %(w/v) PEG 3350
90	0.2 M di-Potassium phosphate		20 %(w/v) PEG 3350
91	0.2 M Ammonium phosphate		20 %(w/v) PEG 3350
92	0.2 M di-Ammonium phosphate		20 %(w/v) PEG 3350
93	0.2 M tri-Lithium citrate		20 %(w/v) PEG 3350
94	0.2 M tri-Sodium citrate		20 %(w/v) PEG 3350
95	0.2 M tri-Potassium citrate		20 %(w/v) PEG 3350
96	0.2 M di-Ammonium citrate		20 %(w/v) PEG 3350

Components of crystallisation screen solutions in PEGs Screen I (Qiagen)

Appendix 5. TmAKR /1PZ0 Sequence Alignment

```
        ....|....| ....|....| ....|....| ....|....| ....|....|
                10         20         30         40         50
TmAKR   MGIPKRKLGE RGPEVSAIGL GCMRMS-FGQ KKLPDRKEMI KLIRTAVELG
1PZ0    --MKKAKLGK SDLQVFPIGL GTNAVGGHNL YPNLNEETGK ELVREAIRNG


        ....|....| ....|....| ....|....| ....|....| ....|....|
                60         70         80         90        100
TmAKR   INFFDTAEVY GPYTNEELVG EALEPFKG-E VVIATKFGFE LYEDGRPGWK
1PZ0    VTMLDTAYIY GIGRSEELIG EVLREFNRED VVIATKAAHR KQGND----F


        ....|....| ....|....| ....|....| ....|....| ....|....|
               110        120        130        140        150
TmAKR   GLNSNPEHIK KAVEGSLRRL RVEAIDILYQ HRVDPNVPIE EVAGAVKELI
1PZ0    VFDNSPDFLK KSVDESLKRL NTDYIDLFYI HFPDEHTPKD EAVNALNEMK


        ....|....| ....|....| ....|....| ....|....| ....|....|
               160        170        180        190        200
TmAKR   EEGKVKHFGL CEASAETIRR AHKVCPVDVV QYEYSMWWRK PEEELLPTCE
1PZ0    KAGKIRSIGV SNFSLEQLKE ANKDGLVDVL QGEYNLLNRE AEKTFFPYTK


        ....|....| ....|....| ....|....| ....|....| ....|....|
               210        220        230        240        250
TmAKR   ELGIGFVAYS PLGKGFLTGA IGENSKFDEE DSRSRIPRFQ KENLRENLAL
1PZ0    EHNISFIPYF PLVSGLLAGK YTEDTTFPEG DLRNEQEHFK GERFKENIRK


        ....|....| ....|....| ....|....| ....|....| ....|....|
               260        270        280        290        300
TmAKR   VELRKTIAER KGATPSQIAL AWLLAQKPWI VPIPGTTKLS HLLENIGGAF
1PZ0    VNKLAPIAEK HNVDIPHIVL AWYLARPEID ILIPGAKRAD QLIDNIKTAD


        ....|....| ....|....| ....|....| ....|
               310        320        330
TmAKR   VELTPEELQE INDALSRIEI KGSRYPEDME KMTYL
1PZ0    VTLSQEDISF IDKLFAPG-- ---------- -----
```

Sequence alignment of TmAKR and 1PZ0, created using Bioedit (Hall *et al*., 1999)

Appendix 6. Stereochemical quality of final TmAKR model
- Main chain parameters
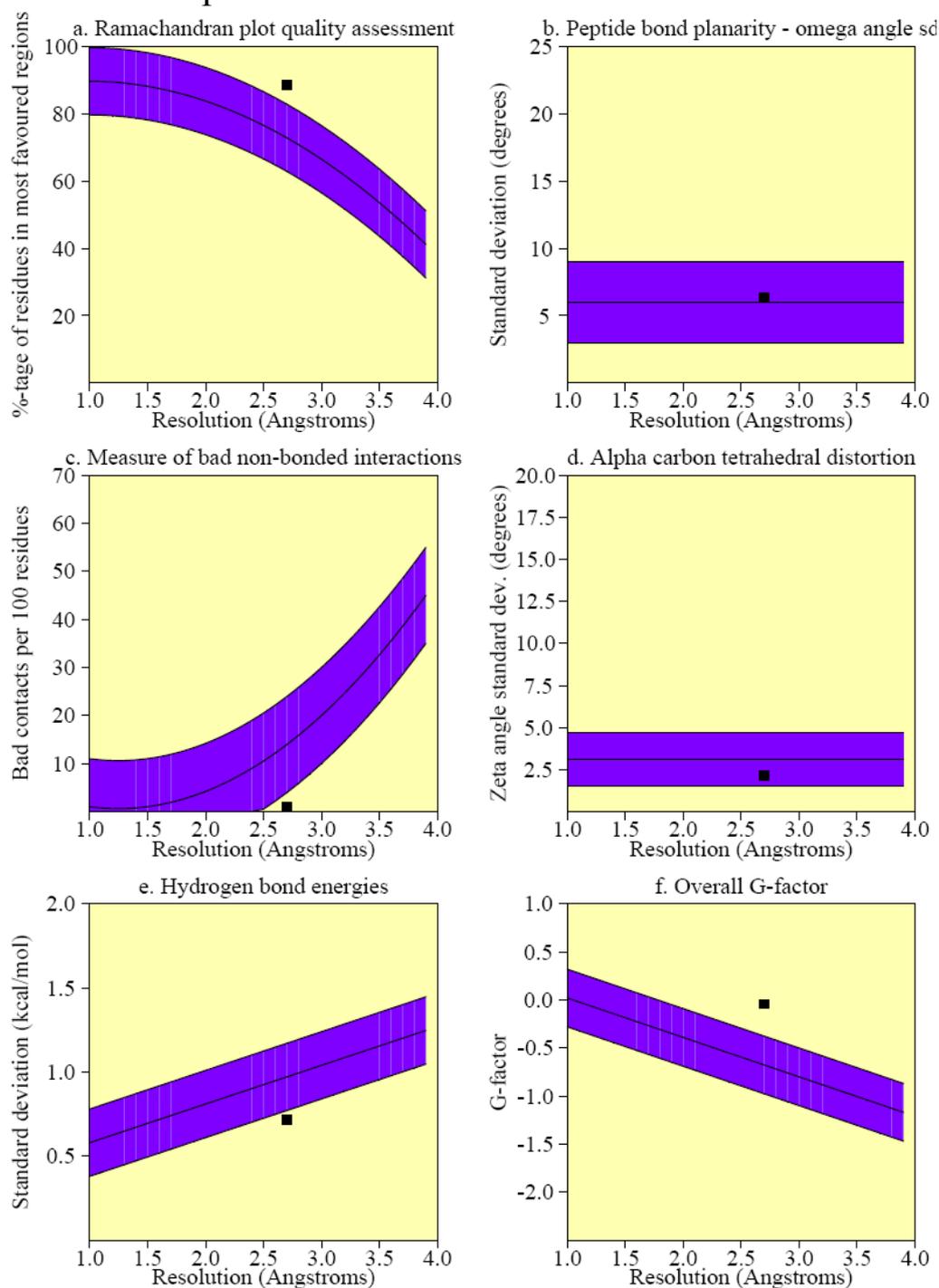
a. Ramachandran plot quality assessment

b. Peptide bond planarity - omega angle sd

c. Measure of bad non-bonded interactions

d. Alpha carbon tetrahedral distortion

e. Hydrogen bond energies

f. Overall G-factor

Plot statistics

Stereochemical parameter	No. of data pts	Parameter value	Comparison values Typical value	Band width	No. of band widths from mean	
a. %-tage residues in A, B, L	481	88.6	72.9	10.0	1.6	BETTER
b. Omega angle st dev	569	6.3	6.0	3.0	0.1	Inside
c. Bad contacts / 100 residues	6	0.9	13.9	10.0	-1.3	BETTER
d. Zeta angle st dev	529	2.2	3.1	1.6	-0.6	Inside
e. H-bond energy st dev	375	0.7	1.0	0.2	-1.3	BETTER
f. Overall G-factor	686	0.0	-0.7	0.3	2.1	BETTER

- Side chain parameters (TmAKR)
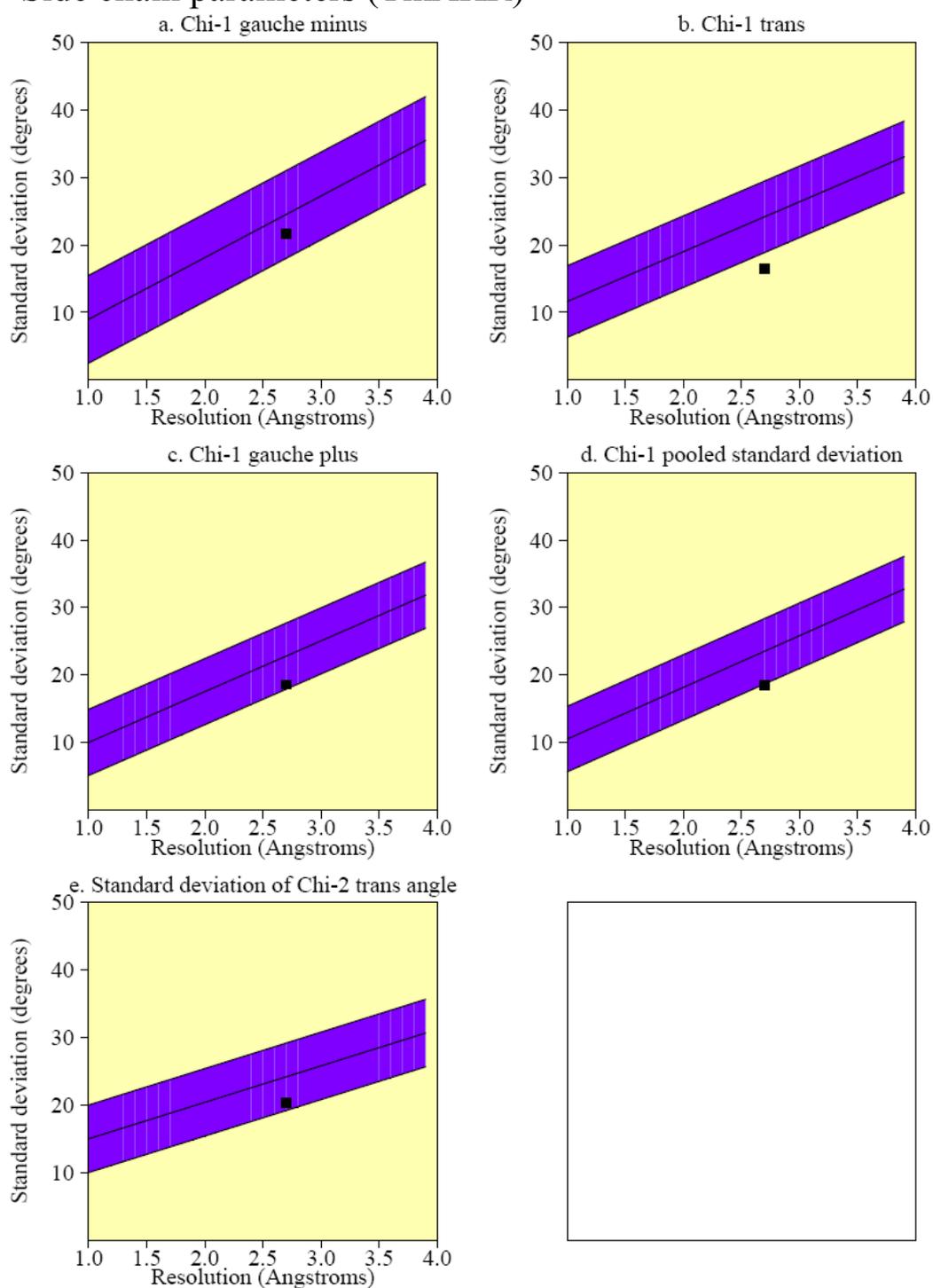
Plot statistics

Stereochemical parameter	No. of data pts	Parameter value	Comparison values Typical value	Band width	No. of band widths from mean	
a. Chi-1 gauche minus st dev	63	21.6	24.5	6.5	-0.4	Inside
b. Chi-1 trans st dev	132	16.4	24.2	5.3	-1.5	BETTER
c. Chi-1 gauche plus st dev	253	18.6	22.8	4.9	-0.9	Inside
d. Chi-1 pooled st dev	448	18.5	23.5	4.8	-1.0	BETTER
e. Chi-2 trans st dev	191	20.3	24.2	5.0	-0.8	Inside

Stereochemical quality of the final TmAKR model. Plots created using Procheck (Laskowski *et al*., 1993)

Appendix 7. Stereochemical quality of Bacterioferritin model

- Main chain parameters
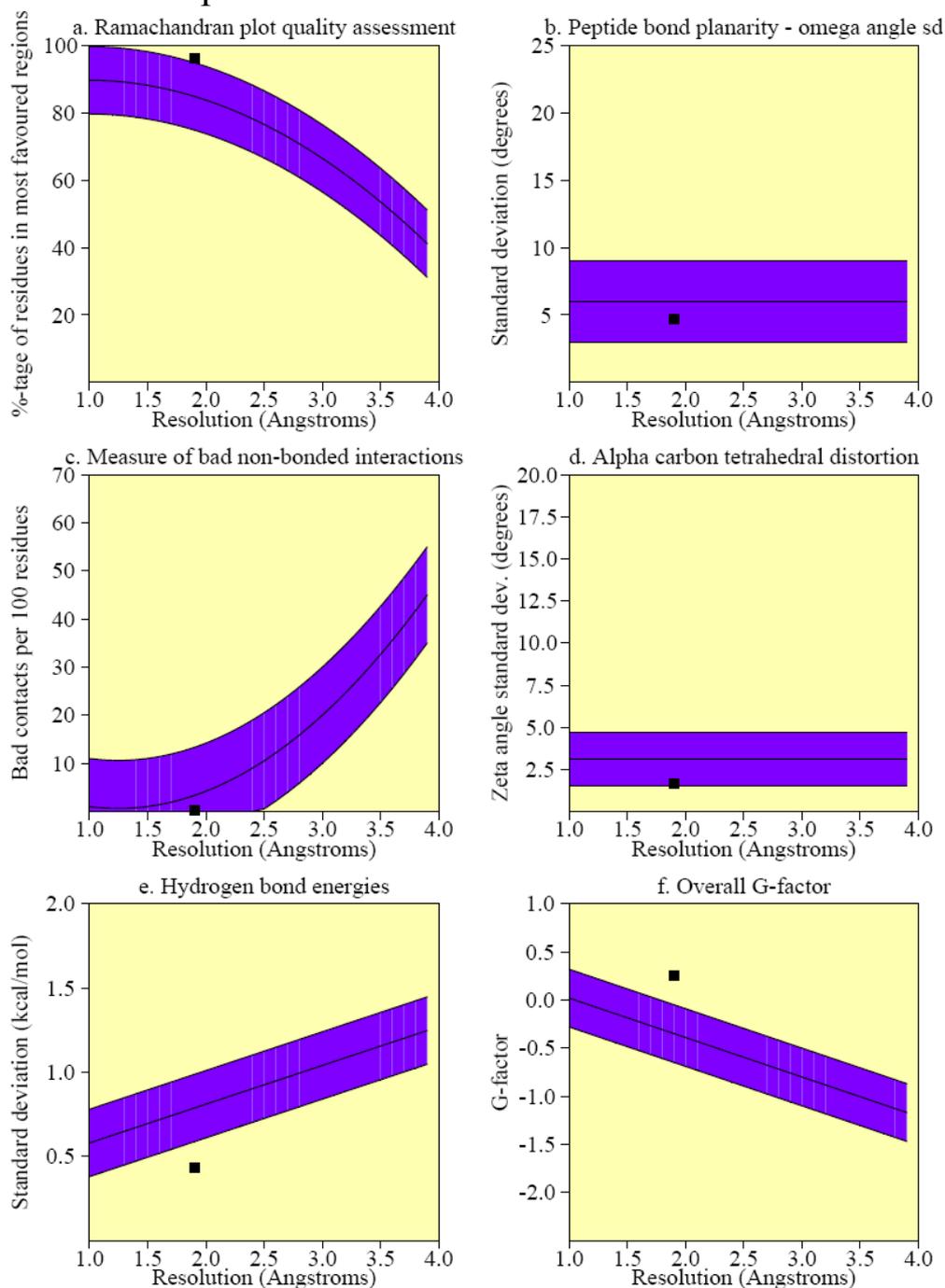
a. Ramachandran plot quality assessment
b. Peptide bond planarity - omega angle sd
c. Measure of bad non-bonded interactions
d. Alpha carbon tetrahedral distortion
e. Hydrogen bond energies
f. Overall G-factor

Plot statistics

Stereochemical parameter	No. of data pts	Parameter value	Comparison values Typical value	Band width	No. of band widths from mean	
a. %-tage residues in A, B, L	1728	96.2	84.9	10.0	1.1	BETTER
b. Omega angle st dev	1872	4.7	6.0	3.0	-0.4	Inside
c. Bad contacts / 100 residues	3	0.2	3.3	10.0	-0.3	Inside
d. Zeta angle st dev	1764	1.7	3.1	1.6	-0.9	Inside
e. H-bond energy st dev	1518	0.4	0.8	0.2	-1.8	BETTER
f. Overall G-factor	1920	0.2	-0.3	0.3	2.0	BETTER

- Side chain parameters (Bacterioferritin)
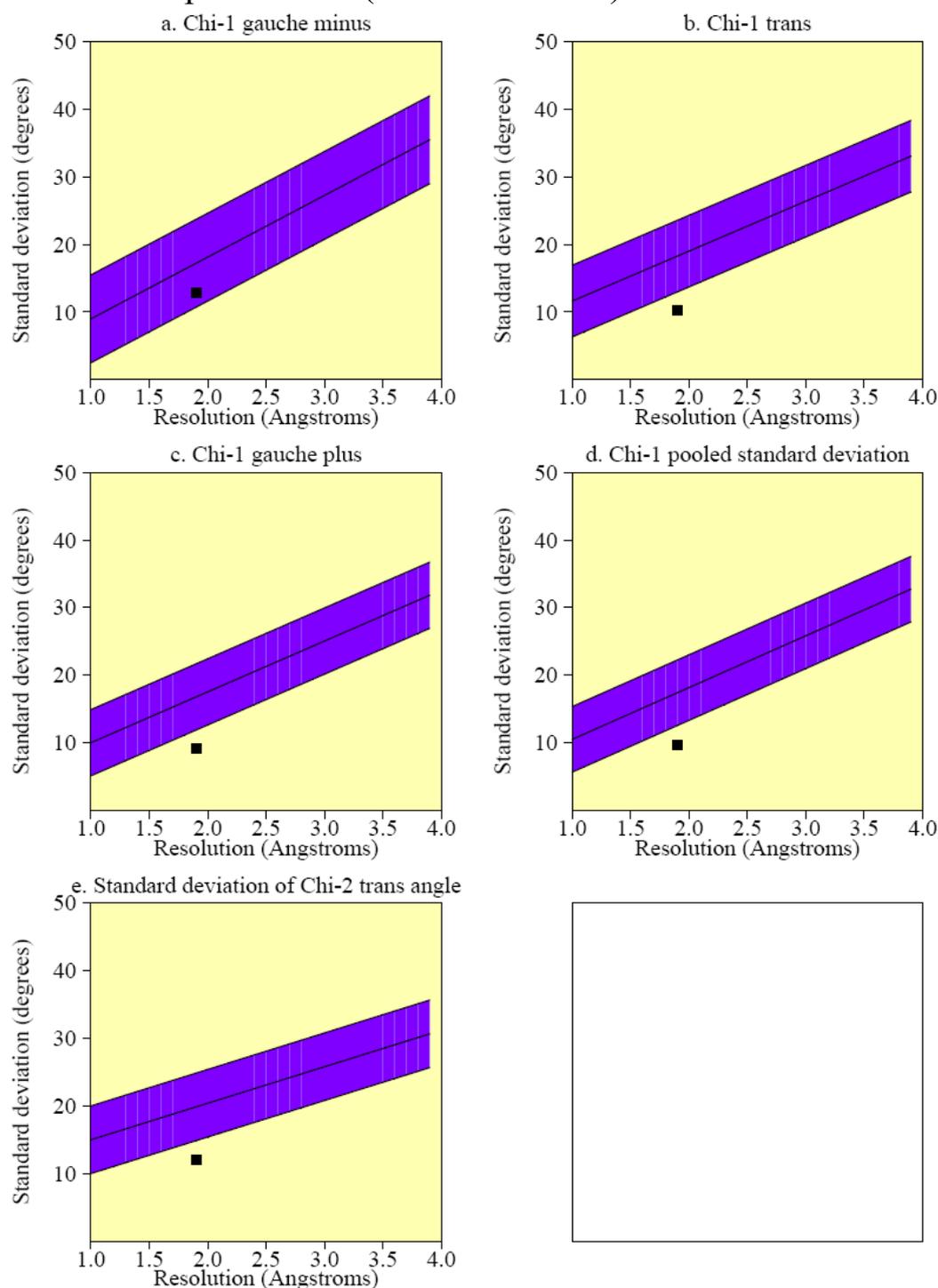
Plot statistics

Stereochemical parameter	No. of data pts	Parameter value	Comparison values Typical value	Band width	No. of band widths from mean	
a. Chi-1 gauche minus st dev	47	12.9	17.2	6.5	-0.7	Inside
b. Chi-1 trans st dev	590	10.3	18.3	5.3	-1.5	BETTER
c. Chi-1 gauche plus st dev	1019	9.2	16.8	4.9	-1.6	BETTER
d. Chi-1 pooled st dev	1656	9.7	17.4	4.8	-1.6	BETTER
e. Chi-2 trans st dev	735	12.1	19.9	5.0	-1.6	BETTER

Stereochemical quality of the final bacterioferritin model. Plots created using Procheck (Laskowski *et al.*, 1993)

Appendix 8. AKR sequence alignment

```
                10        20        30        40        50
       ....|....|....|....|....|....|....|....|....|....|
TmAKR  -MGIPKRKLGERGPEVSAIGLG-CMRMSFGQKKLPDRKEMIKLIRTAVEL
1pz0   ---MKKAKLGKSDLQVFPIGLGTNAVGGHNLYPNLNEETGKELVREAIRN
2alr   --MAASCVLLHTGQKMPLIGLGT---WKSEPGQ------VKAAVKYALSV
2ao0   --MAASCVLLHTGQKMPLIGLGT---WKSEPGQ------VKAAIKYALTV
1pz1   ---MEYTSIADTGIEASRIGLGTWAIGG-TMWGGTDEKTSIETIRAALDQ
1ynp   ---MKKRQLGTSDLHVSELGFGCMSLGTDETKAR-------RIMDEVLEL
1lqa   ---MQYHRIPHSSLEVSTLGLGTMTFGEQNSEAD-----AHAQLDYAVAQ
1mzr   -MANPTVIKLQDGNVMPQLGLGV---WQASNEE------VITAIQKALEV
2he5   MNSKCHCVILNDGNFIPVLGFGTALPLECPKSK------AKELTKIAIDA

                            LOOP 1


                60        70        80        90        100
       ....|....|....|....|....|....|....|....|....|....|
TmAKR  GINFFDTAEVYGPYT-------NEELVGEALEPFK--GEVVIATKFGFEL
1pz0   GVTMLDTAYIYGIGR-------SEELIGEVLREFN-REDVVIATKAAHRK
2alr   GYRHIDCAAIYGN---------EPEIGEALKEDVGPGKAVPREELFVTS
2ao0   GYRHIDCAAIYGN---------ELEIGEALTETVGPGKAVPREELFVTS
1pz1   GITLIDTAPAYGFGQ-------SEEIVGKAIKEYGKRDQVILATKTALDW
1ynp   GINYLDTADLYNQGLN-------EQFVGKALKGRR--QDIILATKVGNRF
1lqa   GINLIDVAEMYPVPPRPETQGLTETYVGNWLAKHGSREKLIIASKVSGPS
1mzr   GYRSIDTAAAYKN---------EEGVGKALKN-----ASVNREELFITT
2he5   GFHHFDSASVYNT---------EDHVGEAIRSKIADG-TVRREDIFYTS

           LOOP 2                              LOOP 3


                110       120       130       140       150
       ....|....|....|....|....|....|....|....|....|....|
TmAKR  YEDGRPGWKGLNSNPEHIKKAVEGSLRRLRVEAIDILYQHRVDPNVP---
1pz0   QGND----FVFDNSPDFLKKSVDESLKRLNTDYIDLFYIHFPDEHTP---
2alr   KLWN------TKHHPEDVEPALRKTLADLQLEYLDLYLMHWPYAFERGDN
2ao0   KLWN------TKHHPEDVEPALRKTLADLQLEYLDLYLMHWPYAFERGDN
1pz1   KNNQ----LFRHANRARIVEEVENSLKRLQTDYIDLYQVHWPDPLVP---
1ynp   EQGKEG--WWWDPSKAYIKEAVKDSLRRLQTDYIDLYQLHGGTIDDP---
1lqa   RNNDKGIRPDQALDRKNIREALHDSLKRLQTDYLDLYQVHWPQRPTNCFG
1mzr   KLWN------DDHK--RPREALLDSLKKLQLDYIDLYLMHWPVPAID---
2he5   KVWC------TSLHPELVRASLERSLQKLQFDYVDLYLIHYPMALKPGEE

           LOOP 3                              LOOP 4


                160       170       180       190       200
       ....|....|....|....|....|....|....|....|....|....|
TmAKR  ----------------IEEVAGAVKELIEEGKVKHFGLCEASAETIRRAH
1pz0   ----------------KDEAVNALNEMKKAGKIRSIGVSNFSLEQLKEAN
2alr   PFPKNADGTICYDSTHYKETWKALEALVAKGLVQALGLSNFNSRQIDDIL
2ao0   PFPKNADGTIRYDATHYKDTWKALEALVAKGLVRALGLSNFSSRQIDDVL
1pz1   ----------------IEETAEVMKELYDAGKIRAIGVSNFSIEQMDTFR
1ynp   ----------------IDETIEAFEELKQEGVIRYYGISSIRPNVIKEYL
1lqa   KLG--YSWTDSAPAVSLLDTLDALAEYQRAGKIRYIGVSNETAFGVMRYL
1mzr   --------------HYVEAWKGMIELQKEGLIKSIGVCNFQIHHLQRLI
2he5   NFPVDEHGKLIFDRVDLCATWEAMEKCKDAGLTKSIGVSNFNYRQLEMIL

                                            LOOP 5
```

206

```
                     210       220       230       240       250
              ....|....|....|....|....|....|....|....|....|....|
      TmAKR    KVCPVD------VVQYEYSMWWRKPEEELLPTCEELGIGFVAYSPLGKGF
      1pz0     KDGLVD------VLQGEYNLLNREAEKTFFPYTKEHNISFIPYFPLVSGL
      2alr     SVASVR------PAVLQVECHPYLAQNELIAHCQARGLEVTAYSPLGSSD
      2ao0     SVASVR------PAVLQVECHPYLAQNELIAHCQARGLEVTAYSPLGSSD
      1pz1     AVAPLH------TIQPPYNLFEREMEESVLPYAKDNKITTLLYGSLCRGL
      1ynp     KRSNIV------SIMMQYSILDRRPEE-WFPLIQEHGVSVVVRGPVARGL
      1lqa     HLADKHDLPRIVTIQNPYSLLNRSFEVGLAEVSQYEGVELLAYSCLGFGT
      1mzr     DETGVT------PVINQIELHPLMQQRQLHAWNATHKIQTESWSPLAQGG
      2he5     NKPGLKYK----PVCNQVECHPYLNQMKLLDFCKSKDIVLVAYGVLGTQR

                                   LOOP 6                LOOP 7


                     260       270       280       290       300
              ....|....|....|....|....|....|....|....|....|....|
      TmAKR    LTGAIGENSKFDEEDSRSRIPRFQKENLRENLALVELRKTIAERKG--AT
      1pz0     LAGKYTEDTTFPEGDLRNEQEHFKGERFKENIRKVNKLAPIAEKHN--VD
      2alr     R-AWRDPDEPVLLEEP-------------------VVLALAEKYG--RS
      2ao0     R-AWRDPNEPVLLEEP-------------------VVQALAEKYN--RS
      1pz1     LTGKMTEEYTFEGDDLRNHDPKFQKPRFKEYLSAVNQLDKLAKTRYG-KS
      1ynp     LSRRPLPEGEGYLNYR----------------YDELKLLRESLPTDRP
      1lqa     LTGKYLNGAKPAGARN-TLFSRFTRYSGEQTQKAVAAYVDIARRHG--LD
      1mzr     K--------GVFDQK--------------------VIRDLADKYG--KT
      2he5     YGGWVDQNSPVLLDEP-------------------VLGSMAKKYN--RT

                   LOOP 7


                     310       320       330       340       350
              ....|....|....|....|....|....|....|....|....|....|
      TmAKR    PSQIALAWLLAQKPWIVPIPGTTKLSHLLENIGGAFVE-LTPEELQEIND
      1pz0     IPHIVLAWYLARPEIDILIPGAKRADQLIDNIKTADVT-LSQEDISFIDK
      2alr     PAQILLRWQVQRK--VICIPKSITPSRILQNIKVFDFT-FSPEEMKQLNA
      2ao0     PAQILLRWQVQRK--VICIPKSVTPSRIPQNIQVFDFT-FSPEEMKQLDA
      1pz1     VIHLAVRWILDQPGADIALWGARKPGQLEALSEITGWT-LNSEDQKDINT
      1ynp     LHELALQYCLAHDVVATVAAGASSIDQVKANVQAVEATPLTAEERQHIQK
      1lqa     PAQMALAFVRRQPFVASTLLGATTMDQLKTNIESLHLE-LSEDVLAEIEA
      1mzr     PAQIVIRWHLDSG--LVVIPKSVTPSRIAENFDVWDFR-LDKDELGEIAK
      2he5     PALIALRYQLQRG--IVVLNTSLKEERIKENMQVFEFQ-LSSEDMKVLDG

                                   LOOP 8


                     360       370       380
              ....|....|....|....|....|....|....|.
      TmAKR    ALSRIEIKGSRYP----------EDMEKMTYL---
      1pz0     LFA-----------------------------
      2alr     LNKNWRYIVPMLTVDGKRVPRDAGHPLYPFNDPY--
      2ao0     LNKNLRFIVPMLTVDGKRVPRDAGHPLYPFNDPY--
      1pz1     ILENTIS-DPVGP----------EFMAPPTREEI-
      1ynp     LAKAAVYEQHRE--------------------
      1lqa     VHQVYTYPAP----------------------
      1mzr     LDQ-------------------GKRLGPDPDQFGG
      2he5     LNRNMRYIPAAIFK--------GHPNWPFLDEY—
```

Protein sequence alignment of TmAKR and closely related structures. Loop regions are highlighted in red