Structural investigations of *E. coli* KDPG aldolase – a carbohydrate synthesising protein

Louise Victoria Buchanan

A thesis submitted for the degree of

Doctor of Philosophy

25th September 2000

Contents

List of Figures

List of Tables

List of Abbreviations

CCD	Charged Coupled Device
CNS	Crystallography and NMR Systems
DNAse	Deoxyribonuclease
DNA	Deoxyribonucleic Acid
DTT	Dithiolthreitol
FBP	Fructose-1,6-bisphosphate
IEF-PAGE	Isoelectric focusing polyacrylamide gel electrophoresis
IPTG	Isopropyl β-D-thiolgalactopyranoside
KDPG	2-keto-3-deoxy-6-phosphogluconate
LB	Luria Broth
NAD(H)	Nicotinamide Adenine Dinucleotide
NAD(P)	Nicotinamide Adenine Dinucleotide (Phosphate)
NCS	Non-Crystallographic symmetry
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
PEG	Poly-Ethylene Glycol
SDS-PAGE	Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis
SeMet	Selenomethionine
TB	Terrific Broth
TRIS	Tris(hydroxymethyl)aminomethane

Declarations

I, Louise Buchanan, hereby certify that this thesis, which is approximately 33000 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree

Date 25/09/00 Signature of candidate

I was admitted as a research student in September 1997 and as a candidate for the degree of Ph.D. in September 1998; the higher study for which this is a record was carried out in the University of St Andrews between 1997 and 2000.

Date 25/09/00 Signature of candidate ...

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Ph.D. in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date ..... Signature of supervis

In submitting this thesis to the University of St. Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and abstract will be published, and that a copy of the work may be made and supplied to any *bona fide* library or research worker.

Date 25/09/00 Signature of candidate ...

Acknowledgements

Firstly thanks to Jim for his endless enthusiasm towards all things protein like and letting me work on the 'odd one out' project. Thanks are also due to our collaborators, the Toone group at Duke University, USA.

A big 'ta' to everyone in the lab, especially Shaggy, your constant interruptions whilst writing this thesis I'm sure I'll look back on fondly; probably in the same way I'll remember the joyous times I've spent at Doom and Grenoble. Quite what you'll do as far as getting the guys in the workshop to do anything for you now I don't think anyone knows!

Immense thanks to my family, don't worry, I promise I won't be coming back to do any more degrees after this one - and yes I know I said that last time as well. Many thanks to the other witches, both to the original ones and those adopted into the coven over the last couple of years.

"I don't see much sense in that" said Rabbit.

"No," said Pooh humbly, "there isn't. But there was going to be when I began it. It's just that something happened to it along the way." (Milne, 1928)

List of Publications

Buchanan, L.V., Wymer, N., Henderson, D., Mehta, N., Botting, C.H., Pocivavsek, L., Fierke, C.A., Toone, E.J. & Naismith, J.H. (2000) "Directed Evolution of a new catalytic site in 2-keto-3-deoxy-6-phosphogluconate aldolase from *Eschericia coli*." Structure folding & design. Accepted.

Buchanan, L.V., Mehta, N., Pocivavsek, S., Niranjanakumari, S., Toone, E.J. & Naismith, J.H. (1999) "Initiating a structural study of 2-keto-3-deoxy-6-phosphogluconate aldolase from *Eschericia coli*." Acta Cryst. D55, 1946-1948.

McMahon, S.A., Leonard, G.A., **Buchanan, L.V.,** Giraud, M-F. & Naismith, J.H. (1999) "Initiating a crystallographic study of UDP-galactopyranose mutase from *Escherichia coli*." Acta Cryst. D55, 399-402.

Abstract

2-keto-3-deoxy-6-phosphogluconate (KDPG) aldolase catalyses the stereospecific re-face addition of a pyruvate enamine to a range of electrophilic substrates, yielding 4-substituted (S)-2-keto-4-hydroxybutyrates. The enzyme has a very limited substrate specificity. Attempts to solve the structure by molecular refinement against P. putida KDPG aldolase were unsuccessful. Thus the structure of E. coli KDPG aldolase was determined to 2.17Å by multiwavelength anomalous dispersion methods using a selenomethionine enriched protein crystal. The enzyme is trimeric, with each monomer being an α/β barrel. A citrate molecule from the crystallisation buffer was found at the active site. A double mutant, K133Q/T161K was identified from directed evolution experiments as having reduced enzyme activity whilst achieving a significantly broader substrate profile than the native. MALDI-TOF spectroscopy confirmed the new lysine to be acting as the Schiff-base forming residue. Modification of substrate specificity is achieved by altering the position of the active site lysine from one β-strand to a neighbouring strand rather than by modification of the substrate recognition site. This new enzyme differs from all other existing aldolases with respect to the location of its active site with regard to secondary structure. The double mutant retains its stereospecificity during aldol addition. The crystal structure of the new enzyme was determined to 2.7Å. These results suggest that the scope of directed evolution is substantially larger than previously envisioned; it is possible to perturb the active site residues themselves as well as the surrounding loops to alter the specificity. The structure of the double mutant shows how catalytic competency is maintained despite spatial re-organisation of the active site with respect to substrate.

Chapter One

Introduction

1.1 Overview

Synthetic organic chemists are increasingly turning to enzymes to make complex stereocentres rather than using entirely traditional methods. However, although enzymes have exquisite regio- and stereo- control they are designed for the compounds nature requires, not those sought after by chemists. The ability to modify or design enzymes to perform custom synthesis has become a Holy Grail of organic chemistry.

The enzyme under investigation in this study, KDPG aldolase, is an example of an enzyme targeted to undergo modification to perform custom synthesis. The introduction discusses the enzyme itself, structural studies of related enzymes and how directed evolution has emerged as a powerful method for the manipulation of enzyme specificity.

Entner Doudoroff Pathway

There are three main pathways responsible for the formation of metabolic precursors which are eventually synthesized into the essential cellular constituents of all living cells. This set of pathways, along with their variations, occurs in all three domains within the living world, strongly suggesting they were established before any divergent evolution between the domains occurred.

The Entner Doudoroff pathway is considered as one of these three original pathways which link into the later stages of glycolysis; the other two being the Embden-Meyerhof-Parnas pathway and the pentose phosphate pathway.

The Entner Doudoroff pathway was first discovered in 1952 in *Pseudomonas saccharophilia*. Entner and Doudoroff provided conclusive evidence for the existence of a novel pathway in *P. saccharophilia* based on the observation that the C1 carbon of glucose was recovered almost entirely as CO_2 (Entner & Doudoroff, 1952). It was found that one molecule of C1 labelled glucose was converted to one molecule each of glyoxal and pyruvate specifically labelled at the carboxyl group. Examination of cell extracts revealed an enzyme capable of cleaving 6-phosphogluconic acid to give pyruvate and glyceraldehyde-3-phosphate. Entner and Doudoroff speculated that 2-keto-3-deoxy-6-phosphogluconic acid was the intermediate formed by the cleavage of a six-carbon compound to two three-carbon compounds. The two steps involved in the reaction were established by separating and purifying the two key enzymes operating in the pathway; 6-phosphogluconate dehydratase and 2-keto-3-deoxy-6-phosphogluconate (KDPG) aldolase. 6-phosphogluconate dehydratase catalyses the degradation of 6-phosphogluconate to KDPG. KDPG aldolase catalyses the aldol cleavage of KDPG to

pyruvate and glyceraldehyde-3-phosphate. It has since been determined that KDPG aldolase is not unique to the Entner Doudoroff pathway whilst 6-phosphogluconate dehydratase is. In a general sense the Entner Doudoroff pathway may be considered to act as a funnel, receiving metabolites derived from the catabolism of a variety of sugar acids, for example, gluconate, glucoronate and galacturonate (Peekhaus & Conway, 1998). The pathway is shown in figure 1.01.

The Embden-Meyerhof-Parnas pathway involves the cleavage of fructose-1,6-bisphosphate by fructose-bisphosphate aldolase to give glyceraldehyde-3-phosphate and dihydroxyacetone phosphate (Conway, 1992). It has been suggested that the common occurrence of modified Entner Doudoroff pathways among saccharolytic archaea, and the absence of the conventional Embden-Meyerhof-Parnas pathway for glycolysis, indicates the Entner Doudoroff pathway is the oldest, and thus original, route of carbohydrate dissimilation (Peekhaus & Conway, 1998).

The primary distinction between the Embden-Meyerhof-Parnas pathway and the Entner Doudoroff pathway lies in the nature of the six-carbon metabolites they use as reactants.

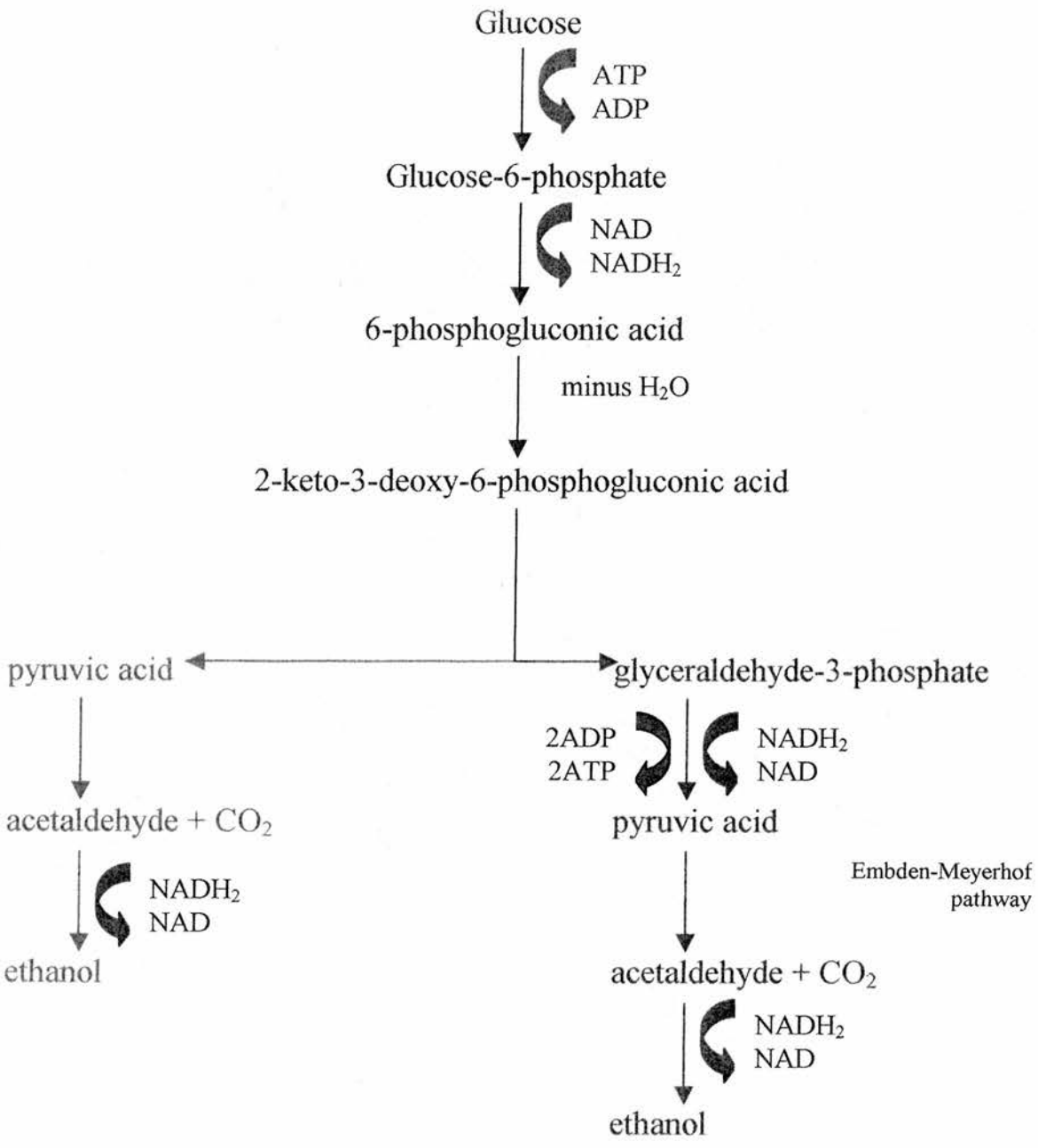


Figure 1.01. The Entner-Doudoroff pathway

As a direct result of the near ubiquity of the Embden-Meyerhof-Parnas pathway among modern bacteria and eucarya the general view taken is this pathway was present in the very earliest organisms (Romano & Conway, 1996). Studies of the amino acid sequences and crystal structures of mainstream Embden-Meyerhof-Parnas pathway enzymes from a wide range of eucaryotic organisms show they are potentially the most highly conserved enzymes known. The complete Embden-Meyerhof-Parnas pathway has not yet been identified in archeon. A modified version of the Entner Doudoroff pathway appears to be the most common method of sugar dissimilation in the saccharolytic archaea studied. It has also been found, from studies of ancient archaea, the Entner Doudoroff pathway can be found to operate in at least two modified schemes using non-phosphorylated intermediates.

In archeon, an incomplete EMP pathway is found; the ED pathway appears to be the most common method of sugar dissimilation. Interestingly in some archeon (*Halobacterium*, *Haloferax* and *Halococcus*) the ED pathway operated with non-phosphorylated substrates (Danson, 1993).

It has been known for some time that an accumulation of phosphorylated intermediates, resulting from the central metabolic pathways, inhibits cell growth although not all are toxic to the cells (Fuhrman *et al.*, 1998). For example, in *E. coli*, *P. aeruginosa* and some Gram-positive bacteria, KDPG is toxic to the cells. In *E. coli* mutations of the *eda* gene are known to lead to inhibition of growth. This has been attributed to an accumulation of KDPG in the cell. Secondary mutations of the gene have been shown to relieve this inhibition and thus prevent the accumulation of KDPG, either by reversing *eda* or by additional mutations in the *edd* gene which is responsible for KDPG

formation. It has been postulated that KDPG toxicity may be the physiological basis of metabolic control in *Z. mobilis* (Fuhrman *et al.*, 1998). Here the genes encoding the glucose transporter and the first three steps of intracellular glucose metabolism are arranged in an operon. The expression of this operon is highly regulated at transcriptional and post-transcriptional levels. It is thought these genetic controls result in an optimal balance of transport and enzyme activities which, in turn, exert a high degree of biochemical control over the flux of glucose to KDPG. When combined such metabolic controls in *Z. mobilis* allows a rapid carbon flux whilst preventing the build-up of toxic intermediates.

Advances in genetic engineering now allow the design of microbes with altered metabolic capabilities. The end products formed from intermediates of central carbon metabolism have even been altered. For example, the *Z. mobilis* genes encoding pyruvate decarboxylase and alcohol dehydrogenase were transformed into *E. coli.* This resulted in a strain of *E. coli* able to convert sugars to ethanol at near theoretical yield, without adversely affecting the physiology of the new strain (Fuhrman *et al.*, 1998).

The Aldol Reaction

The aldol reaction forms carbon-carbon bonds with control of the absolute configurations of the newly formed stereogenic centres (Machajewski & Wong, 2000). This procedure is vital for the industrial synthesis of natural products and pharmaceuticals (Gijsen *et al.*, 1996). The search for methods that predictably transfer chirality efficiently and catalytically by reagent control is one of the most challenging goals in organic synthesis. Several methods have been developed in recent years for the

catalytic asymmetric aldol addition incorporating high density, high efficiency and selectivity.

The aldolase enzymes are a specific group of lyases that catalyse the stereoselective addition of a ketone donor to an aldehyde acceptor. Currently more than thirty aldolases have been identified and these can be classified according to their mechanism (Machajewski & Wong, 2000). Class I aldolases activate the donor by forming a Schiff base as an intermediate in the active site using a catalytic lysine residue. Class II aldolases use a catalytic metal ion, typically Zn^{2+}, as a cofactor in the active site. Each class of aldolase enzyme can be further subdivided into four groups, dependant on the nucleophile used. These four groups are dihydroxyacetone phosphate (DHAP) aldolase, 2-deoxyribose-5-phosphate (DERA) aldolase, glycine dependent and pyruvate / phosphoenol pyruvate dependant aldolases. These will be discussed in turn.

Classification by Nucleophile

The first group of aldolase enzymes use DHAP as the nucleophile, the best-known member of this group being FBP aldolase, see figure 1.02. These aldolases are capable of forming all four diastereoisomers of the 1,3,4-trihydroxy-2-butanone skeleton. This group has been used extensively in the preparation of amino sugars and, to some degree, the precursors of cyclitols (Chou *et al.*, 1995). The aldolase-catalysed reaction of DHAP and pentose or hexose provides easy access to novel long chain sugars which are difficult to obtain from either natural sources or chemical synthesis (Fessner & Walter, 1996; Gijsen *et al.*, 1996).

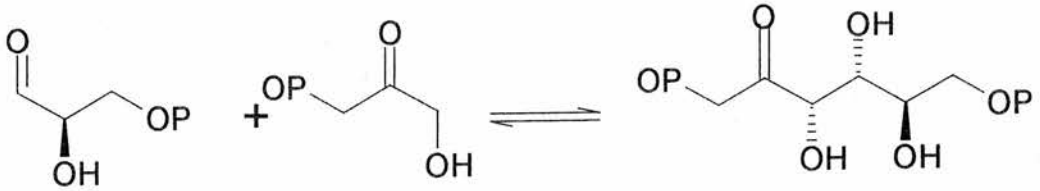
The second group uses acetaldehyde as the nucleophile and consists of a single enzyme DERA, see figure 1.03. This enzyme catalyses the reaction of 2-deoxyribose-5-phosphate from acetaldehyde and D-glyceraldehyde-3-phosphate. To date it is the only known aldolase that performs a cross-aldol reaction of two aldehydes (Machajewski & Wong, 2000). The products of DERA catalysed reactions with acetaldehyde as the donor are also aldehydes; these can then participate as acceptor substrates for a second aldol condensation. DERA can be combined with DHAP-dependent aldolases in a one-pot sequential reaction to give 5-deoxy ketoses with three substituents in the axial position. (Gijsen & Wong, 1995)

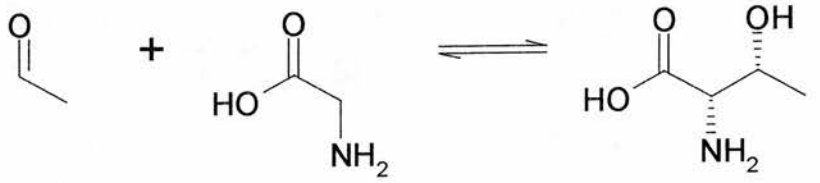
The third group use glycine as the nucleophile, see figure 1.04. These enzymes catalyse the reversible condensation of glycine with an aldehyde acceptor to give a β-hydroxy-α-amino acid. This class has been used extensively for the synthesis of β-hydroxy-α-amino acids, however only a few examples of their use in bond-forming reactions are known. L-Threonine aldolase is known to accept a broad range of substrates (Vassilev *et al.*, 1995). The products are usually a mixture of the *erythro* and *threo* form, with a slight preference for the *erythro* product. The use of hydroxyaldehydes as a substituent gives complex mixtures of products; applying a protecting group to the hydroxyl group solved this problem. The products resulting from L-threonine aldolase catalysed condensations of (benzyloxy) acetaldehyde and (benzyloxy) butanal with glycine have been used in the synthesis of sialyl Le^x mimetics and the immunosuppressant mycestericin D (Shibata *et al.*, 1996; Uchiyama *et al.*, 1995; Wong *et al.*, 1997).

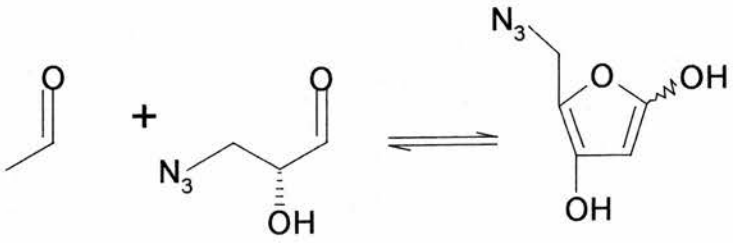
The fourth group, which is the largest in terms of the number of aldolase enzymes encompassed within it, utilises pyruvate or phosphoenol pyruvate as the nucleophile,

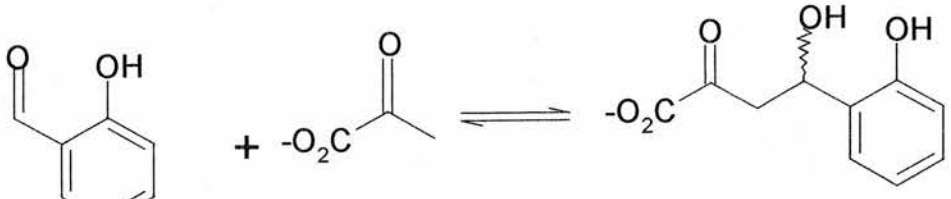
see figure 1.05. They catalyse the formation of a 4-hydroxy-2-ketobutyrate frame. The pyruvate-dependent aldolases for example 2-hydroxybenzalopyruvate aldolase, have a catabolic function *in vivo*; the phosphoenolpyruvate-dependent aldolases are involved in the biosynthesis of keto acids. The conserved four carbon fragments prepared in all pyruvate catalysed reactions are both densely and differentially functionalised.

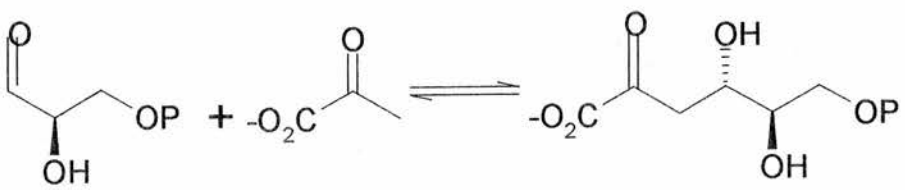
N-acetylneuraminic (NeuAc) aldolase is essentially the most utilised member of the fourth group (Fessner & Walter, 1996; Gijsen *et al.*, 1996). *In vivo* it catalyses the reversible condensation of pyruvate and N-acetylmannosamine to N-acetylneuraminic acid. In the laboratory it is used to prepare modified sialic acids (Lin *et al.*, 1992).











While NeuAc aldolase will accept a range of unnatural electrophilic substrates it does not accept substrates of less than four carbons in length. An added restraint is the enzyme's requirement for polar functionality at position C3 of the electrophile. Under kinetic control NeuAc aldolase catalyses the *si*-face attack of pyruvate. This results in products with the *S*-configuration at the new stereogenic centre. However, most reactions operate under thermodynamic control where the stereochemical outcome of the reaction depends on the configuration of the reactants (Machajewski & Wong, 2000).

2-keto-3-deoxyoctulosonate (KDO) aldolase has been used to prepare both unnatural and natural aldolase sugars. *In vivo* the enzyme catalyses the reversible condensation of pyruvate with D-arabinose to give 3-deoxy-D-*manno*-octulosonate. In a similar manner to NeuAc aldolase, product formation is uniformly under thermodynamic control and product mixtures are frequently encountered. KDO aldolase accepts a variety of polyhydroxylated aldehydes, although lower carbon homologues are accepted albeit with a decrease in reaction rate (Machajewski & Wong, 2000).

KDPG Aldolase Synthetic Utility

Toone *et al.* first published work on the use of KDPG aldolase (figure 1.06) as an enzyme for synthetic organic chemistry in 1992 (Allen *et al.*, 1992). This initial study showed the enzyme could accept unnatural electrophiles at synthetically useful rates. Initially ten aldehydes were tested, the rates of reaction being reported relative to that for D-glyceraldehyde, itself accepted at approximately 0.8% the rate of natural substrate. All of the substrates were utilised at much lower rates than the natural substrate; a phenomenon also noted for DERA (Barbas *et al.*, 1990). The reactions did

proceed however with the expected stereochemistry. KDPG aldolase was found not to utilise simple aliphatic aldehydes or benzaldehyde.

A second study compared KDPG aldolase from three bacterial sources; *E. coli*, *P. putida* and *Z. mobilis* (Shelton *et al.*, 1996). The enzyme from *Azobacter vinelandii* was studied previously (Taha & Deits, 1994). The enzymes from each source were investigated for enzyme stability, ease of isolation, organic solvent stability, pH optima, substrate specificity and suitability for preparative scale reactions. A summary of the results of this study may be found in table 1.01.

Characteristic	*Z. mobilis*	*P. putida*	*E. coli*
Ease of isolation	Bad	Good	Very good
Enzyme stability	Bad	Very good	Good
Activity in organic solvents	Very good	Very good	Very good
pH optima	Good	Good	Very good
Substrate specificity	Fair	Good	Good
Usefulness for preparative scale reactions	Bad	Very good	Very good

Table 1.01 Comparison of KDPG aldolase from three sources

All three enzymes show substrate specificity which is broad enough to be useful in organic synthesis. However, in all three the removal of the phosphate group from the glyceraldehyde-3-phosphate has a pronounced effect on the rate of reaction. D-glyceraldehyde is converted to 2-keto-3-deoxygluconate at approximately 1% of the rate that D-glyceraldehyde-3-phosphate is converted to 2-keto-3-deoxy-6-phosphogluconate. The negative charge on C3 of the electrophile appears to be optimal; the charge may also be located at C2 or C4. Thus, two good unnatural electrophiles are D-erythrose-4-

phosphate and glyoxylate which are used by all three aldolases at greater than 25% the rate of D-glyceraldehyde-3-phosphate.

Overall, despite many similarities in substrate specificity, the three enzymes show several significant differences. *P. putida* and *E. coli* have very similar and relatively restrictive substrate specificity; *Z. mobilis* shows less selectivity and accepts a broader range of substrates. The absolute requirements for reaction to occur were found to be a negatively charged residue at position 2, 3 or 4 of the electrophile, or substituents capable of acting as hydrogen bond acceptors, and minimal steric hindrance at C3. In addition both *P. putida* and *E. coli* require the D-configuration at C2 of the aldehyde substrate.

To establish the level of asymmetric induction available, preparative scale reactions were carried out using D-glyceraldehyde, D-erythrose and D-lactaldehyde as electrophilic substrates. The diasterioselectivity of the reactions was determined by ^{13}C NMR of the upfield C3 methylene signal. *P. putida* and *E. coli* gave a diasteriomeric excess of greater than 97%; for *Z. mobilis* the result was only 50%.

In retro aldol reactions KDPG aldolase is specific for the open chain form of 2-keto-3-deoxy-6-phosphogluconate (approximately 9% in aqueous solution). This is in contrast to NeuAc aldolase and KDO aldolase which both bind cyclic substrates (Kim *et al.*, 1988; Kragl *et al.*, 1994; Lin *et al.*, 1992; Sugai *et al.*, 1993). Specificity for the open chain form in this reaction type has profound implications for both the reversibility of the reaction and the mechanism of diasterioselectivity. For example, 3-deoxy-D-erythro-hexulosonate cannot be cleaved to pyruvate and glyceraldehyde even in the

presence of lactate dehydrogenase. In such a case the observed diasterioselectivity in the aldol condensation is purely kinetic.

It was proposed that products which exist in the open chain form, at concentrations nearing or greater than the K_m for retro aldol cleavage, could, in principle, be formed under either kinetic or thermodynamic control. Enzymatic condensation of pyruvate with enantiomerically pure D-lactaldehyde could give the two equienergetic diastereomers. That only one diastereoisomer is produced implies that the enzymatic aldol condensation remains entirely under kinetic control despite the reversibility of the reaction. NeuAc aldolase and KDO aldolases, in contrast, operate under either kinetic or thermodynamic control, depending on stereochemical configuration of the substrates.

Phosphorylation Effect

A noted drawback of KDPG aldolase, and pyruvate aldolases in general, is the relatively low rate of reactivity with unnatural electrophilic substrates compared to natural ones (Cotterill *et al.*, 1998). The most pronounced loss of reactivity involves the removal of the phosphate group of the natural electrophile. As mentioned D-glyceraldehyde reacts at only 1% the rate of D-glyceraldehyde-3-phosphate. Most unnatural electrophiles are accepted at, or near, the rate of D-glyceraldehyde. Whilst the relative activities of the enzymes are low, the absolute specific activity remains sufficiently high for preparative scale organic synthesis.

It was proposed that the phosphorylation of electrophilic substrates might improve the rate of their acceptance by KDPG aldolase. It was also thought phosphorylation of the terminal hydroxy group would facilitate conversion of the aldol adduct to a range of

biologically relevant products. Experimentally the effect of phosphorylation on the rate of enzymatic reaction was found to vary widely and no systematic or universal variation in reaction rate in response to phosphorylation was identified. The pattern of rate enhancement / depression was found to be identical for KDPG aldolase from both *E. coli* and *Z. mobilis*. The *Z. mobilis* aldolase accepted all unnatural substrates at significantly greater rates than the *E. coli* one. This is consistent with previously observed results that the *Z. mobilis* enzyme is the most promiscuous of all of the KDPG aldolase enzymes examined.

In conclusion, it was found that even in the absence of a significant rate enhancement phosphorylation of the electrophilic hydroxy group of unnatural substrates provided unique differential production of product. In this capacity the phosphate group effectively blocks cyclisation of the adduct through this hydroxy group, providing the product exclusively in the open-chain form. Overall, while the phosphate functions as a useful protecting group in enzymatic aldol addition, phosphorylation produces a modest and unpredictable effect on reaction.

α/β Folds

There are three main classes of α/β proteins. The most dominant has a core of twisted parallel β strands arranged closely together, in a form reminiscent of the staves of a barrel. The α helices that connect the parallel β strands are located on the outside of this barrel. This domain structure is known as the TIM barrel after the enzyme, triose phosphate isomerase, whose structure was determined in 1972, shown in figure 1.07 (Banner *et al.*, 1975). This is a very common protein fold.

The second class contains an open twisted β sheet surrounded by α helices on both sides. A typical example is a nucleotide-binding domain, known as the Rossmann fold, shown in figure 1.07, which was first identified in lactate dehydrogenase in 1970. The third class is formed by amino acid sequences containing repeat regions of a specific pattern of leucine residues that form α helices and β strands. The β strands form a curved parallel β sheet with the α helices on the outside. One of the members of this class, a ribonuclease inhibitor, is shaped like a horseshoe; this class is known as the horseshoe fold, shown in figure 1.07.

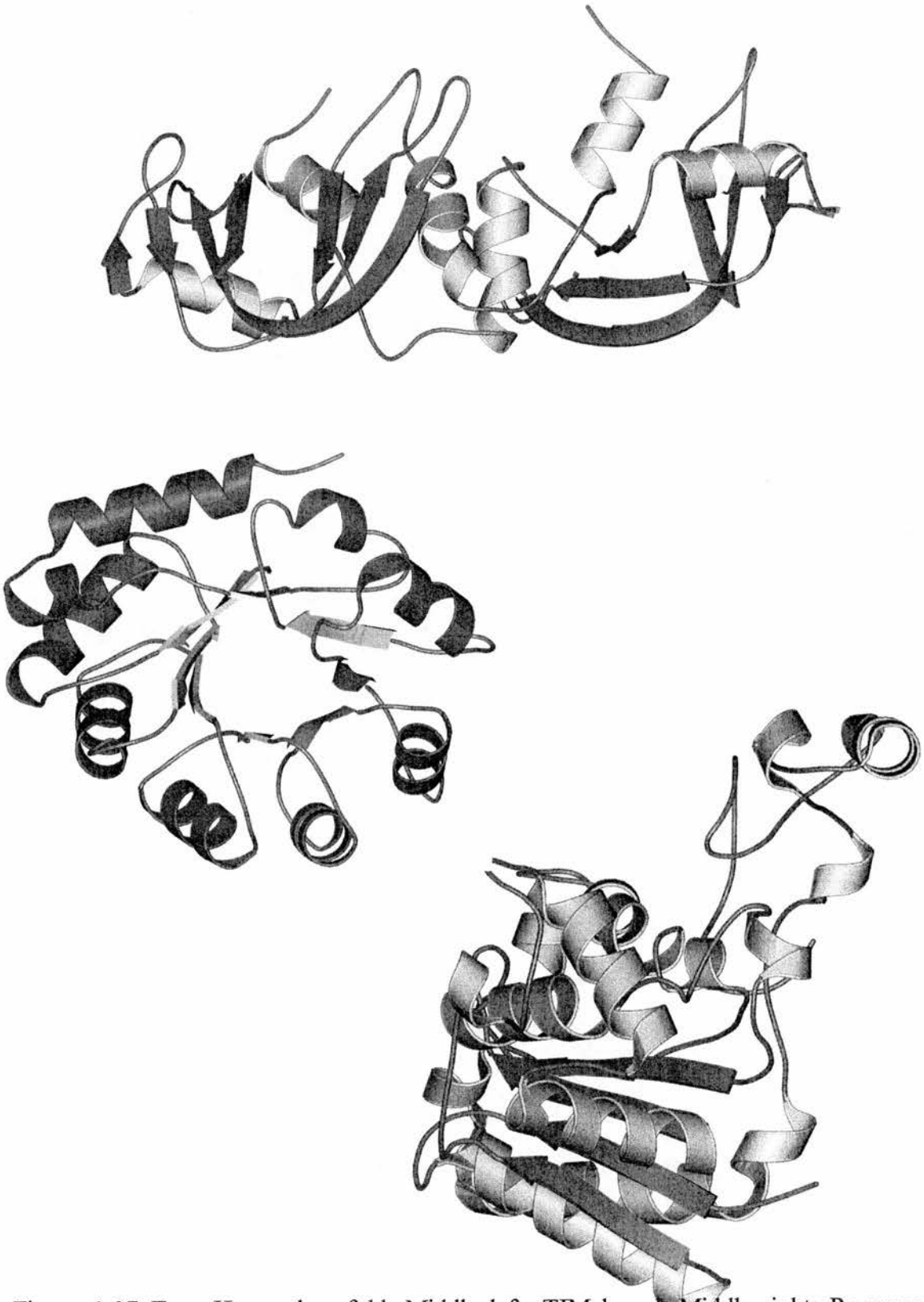


Figure 1.07 Top: Horse shoe fold. Middle left: TIM barrel. Middle right: Rossmann fold.

The TIM-barrel fold

The $(\alpha/\beta)_8$ barrel is one of the most common of all the domain structures. A minimum of 200 residues are necessary to form this arrangement which has been found in many different proteins. The majority of these proteins are enzymes with almost no consensus sequence encoding for the fold. Superpositioning the structures of TIM-barrel proteins shows 160 residues to be structurally equivalent; these are the residues that form the α helices and β strands. The remaining residues form the loop regions that connect these strands and helices. These loops have quite different lengths and conformations. It is accepted that the strands and helices form the very stable structural framework while the loops control the substrate specificity.

In $(\alpha/\beta)_8$ barrels the hydrophobic side chains of the β sheets are packed against the hydrophobic side chains of the α helices. The α helices are anti-parallel and adjacent to the β sheets which they connect to on the barrel. The barrel is therefore provided with a shell of hydrophobic residues originating from both the strands and the helices. It is also seen that the side chains of the residues from the β sheet involved in this shell formation come from every second residue of the sheet. Thus the other side chains from the β sheet point inside the barrel to form a hydrophobic core.

Evolution of New Enzymes

The question which has intrigued researchers for many years' concerns how the very different attributes of enzymes with the TIM barrel fold evolved. Did they arise by divergent evolution from a pre-existing set of parent enzymes or were they formed by consequence of the very stable structural framework of the fold? Recent studies have provided evidence for the former from studies of α/β barrel enzymes in a rare metabolic

pathway, notably the conversion of mandelate to benzoate, which is present in only a few *pseudomona* species.

Aldolase Structures & Mechanism

FBP aldolase is a key enzyme in cellular metabolism, catalysing the reversible aldol condensation / cleavage reaction between glyceraldehyde-3-phosphate and DHAP to yield FBP. The two classes of FBP aldolase are distinguished by their structural and mechanistic properties. The Class I aldolases utilise an active site lysine to form a Schiff base with the substrate as part of the reaction mechanism (Verlinde & Quigley, 1999). They are predominantly found in higher life forms; animals, plants, protozoa and algae. The Class II aldolases are metallo-enzymes which use a bivalent metal ion in catalysis. These aldolases are found in unicellular organisms; bacteria, yeast, fungi and blue-green algae (Hall *et al.*, 1999).

Almost all organisms, therefore, contain only one class of FBP aldolase. However, *E. coli* is one of the few organisms which express both types of FBP aldolase. This duplication is thought to represent a catalytic redundancy that has been eliminated from most organisms through evolution. Crystal structures of both classes of FBP aldolase from a variety of sources have been solved. They adopt the α/β barrel fold first identified as the TIM barrel. From direct sequence comparisons it has been shown that approximately 15% sequence identity exists between Class I and Class II FBP aldolases.

Class I Aldolases

FBP aldolase from rabbit skeletal muscle aldolase was solved to 1.9Å (Blom & Sygusch, 1996). Examination of the active site showed DHAP to be bound in two of the four subunits. The presence of the DHAP bound to the aldolase was a consequence of purification and crystallisation and as such was not intentional. Binding of the DHAP, which is present as the free keto form and thus as its most reactive, involves extensive hydrogen bonding of its C_1-phosphate oxyanion, C_2-carbonyl oxygen and C_3-hydroxyl. It was also noted that the almost perpendicular orientation of the Schiff-base forming residue, Lys 229, with respect to the C_2-carbonyl bonding plane is sterically consistent with facile tetrahedral carbinolamine intermediate formation at the C_2-carbonyl carbon. Lys229 is 3.1Å from the DHAP C2. It was proposed that the conformation of the C-terminal region of the enzyme mediates reactant attachment and would control Schiff-base formation and thus the activity of the enzyme.

To investigate the active site FBP aldolase crystals from human muscle were soaked with FBP in the expectation an equilibrium would occur between the substrate and product (Dalby *et al.*, 1999). Examination of the initial difference map, from data to 2.8Å, showed a high proportion of substrate to be bound. The orientation of the substrate was found to be significantly different from that proposed by previously (Gamblin *et al.*, 1991). However it was in agreement with the binding of DHAP to rabbit muscle FBP aldolase.

FBP aldolase from rabbit muscle may be regarded as the classical reference catalytic enzyme because it has been tested with hundreds of aldehydes as substrate analogues in place of the natural substrate. Its main limitations, with respect to synthetic chemistry

applications, remain its relatively low stability under standard reaction conditions and a half-life measurable in days in buffered solutions (Fessner, 1998).

An important study in the elucidation of the mechanism of catalysis involved the enzyme Transaldolase B, D-sedoheptulose-7-phosphate:D-glyceraldehyde-3-phosphate di-hydroxyacetone transferase (Jia *et al.*, 1996). This aldolase originates from the non-oxidative branch of the pentose phosphate pathway and catalyses the reversible transfer of a dihydroxyacetone moiety, derived from fructose-6-phosphate, to erythrose-4-phosphate, yielding sedoheptulose-7-phosphate and glyceraldehyde-3-phosphate. Catalysis occurs via a Schiff base between Lys132 and the dihydroxyacetone moiety; a typical Class I mechanism. The enzyme is found in both prokaryotic and eukaryotic species. Sequence alignment has shown there is 53% identity between the enzymes from *E. coli* and *S. cerivisiae* and 55% between *E. coli* and human transaldolase.

The active site is located at the C-terminal end of the β strands. The walls of the active site cavity are formed by the loops connecting the β strands with the α helices, akin to other α/β barrel enzymes. In this structure this cavity is accessible from the bulk solvent despite helix αF running across half of the funnel-shaped opening of the cavity. At the bottom of this cleft is located the ε-nitrogen of Lys132, which forms the Schiff base.

Comparisons of transaldolase B with other members of the Class I aldolase family which have been solved crystallographically have shown their overall topologies to be similar. Superimposing transaldolase and FBP aldolase, chosen due to the similarity of the reaction they catalyse, by their secondary structural elements, β1 with β1, β2 with β2 and so on, gives interesting results. The alignment results in 113 equivalent Cα's

with an r.m.s. deviation of 2.53Å. However, the two Schiff base forming lysine residues, Lys132 for transaldolase and Lys229 for FBP, are not aligned.

An alternative technique to align the two enzymes is to align the β strands carrying the Schiff base forming lysine residues; β4 of transaldolase with β6 of FBP aldolase. This method results in 146 equivalent Cα's with an r.m.s. of 2.17Å. Overall the two enzymes are structurally very similar despite the apparent lack of significant overall sequence homology. From the high degree of overall structural similarity, together with the conservation of some active site residues and the phosphate binding site, it is probable that both enzyme families have evolved from a common ancestor. However, optimal alignment of the active site residues in FBP and transaldolase requires a circular permutation of the secondary structural elements of the transaldolase. As an aside this exemplifies the internal rotation symmetry present in TIM barrels. This internal symmetry often confounds attempts at structure determination using them as search models for molecular replacement techniques.

It was proposed that such a permutation in an ancestral aldolase would have occurred during evolution when the first two β strands of this pre-aldolase were moved to the C-terminus. An important point to note is that in N-acetylneuraminate lyase, dihydrodipicolinate synthase and KDPG aldolase, as well as FBP aldolase, the Schiff base forming lysine is located on strand β6. The similarity in the chemistry and substrates between FBP aldolase and transaldolase implies these two enzymes are more closely related to each other than the other family members. It was concluded that the common ancestor of the Class I aldolase family had the Schiff base forming lysine on strand β6.

Naturally occurring circular permutations of proteins within a single domain have been found in other examples, notably the saposin domains and the β-gluconases. However, it may also be that convergent evolution imposed by the chemical constraints of the reaction to be catalysed is the driving force. The high structural similarity found between the structures indicates a circular permutation has occurred (Jia *et al.*, 1996).

The structure of transaldolase B complexed with the reduced Schiff base intermediate was determined to provide an insight into the mechanism of action of Class I aldolases (Jia *et al.*, 1997). This was the first crystallographic description of the Schiff base intermediate at the active site of a Class I aldolase. They found the overall structures of the native and native plus intermediate to be very similar, thus no notable conformational movement occurs upon substrate binding. A comparison of the two gave an r.m.s. deviation for the 316 Cα's of 0.13Å. Based on combined biochemical and structural data a reaction mechanism was derived. The main features of this mechanism included Lys132 as the Schiff base forming residue and two acidic groups involved in proton transfer during the reaction. A catalytic water molecule was also found to be present in the active site.

Class II Aldolase Structures

Less is known about Class II aldolases than their Class I counterparts. They are more stable than the Class I enzymes and offer novel opportunities in bio-transformation chemistry, particularly if an understanding of their mechanism and selectivity can be obtained (Cooper *et al.*, 1996). As this class of aldolases are not found in mammals it is possible they represent a target for anti-bacterial agents. A thorough understanding of

the enzymes has implications for a structure based inhibitor programme. Class II FBP aldolases have an absolute requirement for a divalent cation, typically Zn^{2+}, and are activated by monovalent cations (Hall *et al.*, 1999).

The structure of Class II FBP aldolase from *E. coli* was solved to 2.5Å in 1996 (Cooper *et al.*, 1996). The $(\alpha/\beta)_8$ barrel core was found to be circular in cross section. The active site of FBP aldolase from *E. coli* has a novel bimetallic binding site; one metal ion is buried with the other is on the surface of the active site. While the identity of the latter metal ion is known to be a catalytic zinc ion, the identity of the former was ambiguous and thought to potentially be potassium, zinc or even a mixture of the two.

Cooper *et al.* superimposed this structure with that of the Class I enzyme from *Drosophila*. Alignment using the central portions of the β strands 1 to 8 of the Class I enzyme with those of the Class II gave an r.m.s. deviation of 2.3Å for 124 Cα atoms. The key residues in each structure are located close to each other when they are aligned by structural elements. In fact the N-ε atoms of the Schiff base forming lysine, Lys229, and the suspected orientating lysine, Lys146, are found within 1Å of the N-ε2 atoms of two of the catalytic zinc liganding residues, His110 and His264. Alternative overlaps, varying the strand alignments, decrease the number of Cα atoms that overlap within 2.5Å and do not align residues implicated in the enzyme mechanism. This suggests the circular permutation thought to have occurred in the Class I aldolase family, (Jia *et al.*, 1996), is not applicable to the Class II family; presuming they have evolved from a common ancestor.

Class II FBP aldolase, from *E. coli*, complexed with phosphoglycolohydroxamate (PGH) was solved to 2Å (Hall *et al.*, 1999). In previous studies of Class II enzymes mutually exclusive Zn^{2+} sites with approximately 50% occupancy were identified. In this structure one Zn^{2+}, the buried one, is absent while the surface one is fully occupied and well ordered. This is proposed as the catalytic zinc ion. An examination of the aldolase-PGH interaction showed the enzyme uses a combination of structural elements to construct the binding site including the loop regions of the β6-loop-α8, β7-loop-α9, the C-terminus end of β8 and the N-terminus end of α10. This sector of the α/β barrel is associated with binding phosphate groups and as such contains features often associated with protein-phosphate interactions. It has been shown that approximately 50% of phosphate groups bind near the amino-terminal of an α-helix.

In Class I aldolases formation of the Schiff base requires the catalytic lysine to be deprotonated to allow it to act as a nucleophile to attack the substrate's carbonyl function. The electron withdrawing ability of the resulting iminium ion facilitates the aldol condensation / cleavage reaction. In Class II aldolases a similar electron withdrawing function is performed by the active site metal ion; this is then thought to stabilise an enolate reaction intermediate (Plater *et al.*, 1999).

5-Aminolaevulinate Dehydratase (ALAD) catalyses the condensation of two 5-aminolaevulinic acid to give pyrrole porphobilinogen. Four porphobilinogen molecules are then condensed to form a linear tetrapyrrole, which is cyclised and rearranged to give the first macrocyclic tetrapyrrole of the pathway for tetrapyrrole biosynthesis. The three dimensional structure of ALAD is of interest primarily because the asymmetric Knorr-type condensation reaction catalysed by the enzyme combines mechanistic

features previously associated with both Class I and Class II aldolase enzymes (Erskine *et al.*, 1997).

ALAD encompasses a TIM barrel fold with the Schiff-base forming lysine, Lys263, located on strand β6; an aldolase Class I characteristic. However, there is a zinc ion bound to the protein; an aldolase Class II characteristic. It remains unclear how this enzyme functions.

Structural Studies on P. putida KDPG Aldolase

The three-dimensional structure of the Class I KDPG aldolase from *P. putida* was solved to 3.5Å in 1976 (Mavridis & Tulinsky, 1976). X-ray data showed the enzyme to be composed of three identical subunits, providing evidence for the first well documented case of an enzyme existing in nature as a trimer (Vandlen *et al.*, 1973). The rare occurrence of trimeric proteins in nature has led to the theory there may be fundamental constraints operating against the evolutionary survival of odd-numbered oligomers when compared with the much more abundant even-numbered ones. Investigations showed the protein to consist of assemblages of four trimers in the crystal, with the subunits related to each other by a crystallographic three-fold rotation axis. The overall profile of the oligomeric structure, when viewed down the three-fold axis, resembled a ship propeller.

At 3.5Å the tracing of the polypeptide chain was greatly facilitated by the absence of disulfide bridges, and thus main-chain branching, in the density maps. The authors do note that due to the limitation of only having medium resolution data assumptions had to be made regarding the direction of the main chain at certain points. These sections

were generally where the density was not connected and / or more than one possibility existed as to which density path to follow. However, it is stated these regions are minimal and most of the backbone could be traced unambiguously. The two specific regions where interpretation of the maps was not elementary were traced using the past experience and intuition of the authors (Mavridis & Tulinsky, 1976).

At 3.5Å *P. putida* KDPG aldolase was described as containing nine α-helices and a small number of parallel β-sheets, although these were not readily organisable into a describable overall pattern. In 1979 Richardson examined the C-α co-ordinates (Richardson, 1979). From this examination she found them to be organisable into a rose pattern, very similar to the end-on view of the singly wound parallel β-barrels of the first domain of pyruvate kinase (Levine *et al.*, 1978) and triosephosphate isomerase (Banner *et al.*, 1975).

Richardson suggested that two reconnections of the backbone of *P. putida* KDPG aldolase would provide the central portions of the 'missing' β-sheets; these adjustments increasing the proteins' similarity to the folds found in pyruvate kinase and triosephosphate isomerase. These alterations transformed the structure into a singly wound, eight stranded, parallel β barrel. Discussions with Tulinsky led him to re-examine the maps at these points; no density was evident in these areas and any new backbone tracing would have had to cross clear gaps in the density (Richardson, 1979). However, in other crystal structures in the 2.7-3.5Å resolution range, the medium to low resolution range, during this stage in the evolution of protein crystallography the most common mistake was the incorrect tracing of false connectivities running parallel between a pair of true β-strands. At this time the overall organisation and relative

placement of the secondary structure elements were dependable but the details of backbone connectivity were notoriously unreliable for this resolution. At this point the debate regarding the true conformation of the protein was put on hold, with the belief the detailed connectivity would turn out to be that of a singly wound parallel β-barrel.

In 1982 Tulinsky published, but did not deposit in the PBD, *P. putida* KDPG aldolase refined to 2.8Å (Mavridis *et al.*, 1982). At this higher resolution, and after the amino acid sequence was determined in 1980 (Suzuki & Willis, 1980), all 225 residues were found; in the 3.5Å structure only 173 residues were located. The structure was initially re-examined because of the significant work done between 1976 and 1982 involving glycolytic pathways. Much progress was made in delineating the structure and function of the glycolytic enzymes, however examination of aldolase enzymes was found to be lacking. Thus, the original investigators decided to contribute to the 'most important inter-related general study of glycolytic enzymes' by extending the resolution of the structure of *P. putida* KDPG aldolase to 2.8Å and thus be able to include a description of side chain orientation and stereochemistry.

As mentioned above, all of the 225 residues belonging to the protein were found. Essentially the missing residues were not found to be located as an appendage to the 173 residues originally found at 3.5Å but in the interior of the molecule. Examination of the structure found that the complete interior folds regularly, in a manner strikingly similar to triosephosphate isomerase. In this protein fold the inner cylinder consists of eight strands of parallel β-sheets linked by helical segments and loops of non-regular backbone (Mavridis *et al.*, 1982).

As mentioned previously Richardson proposed an $(\alpha/\beta)_8$ barrel organization of the protein on examination of the C-α backbone at 3.5Å (Richardson, 1979). It was found that four chain reconnections could introduce four additional β-strands, resulting in a change in protein folding. Richardson was correct in so far as the $(\alpha/\beta)_8$ barrel arrangement of the protein is concerned, however because the additional 50 residues located in the 2.8Å data were found in the interior of the protein, the details of the barrel are different. Thus, whilst the reconnections suggested apply only in a topological sense, Richardson's insight was remarkable.

Directed Evolution

Directed evolution exploits the methodology nature has fine-tuned over millions of years for altering an enzymes function and places it at the mercy of scientists. The ability to create protein's *de novo* with novel functions and properties is a Holy Grail of protein engineering (Nixon & Firestine, 2000). Directed evolution involves making random mutations in the gene encoding an enzyme of interest and does not require any structural or mechanistic information about the protein or the reaction it performs (Jaeger & Reetz, 2000).

Protein Sequence Space

Sequence space is a useful concept applicable to the targeting of enzymes for directed evolution (Reetz & Jaeger, 2000; Steipe, 1999). Every sequence is considered to be represented by a point in space; their dimensions of space correspond to their positions in the sequence. Normally they equal the sequence length but may be larger when gaps are required. A walk in sequence space may be considered to be a series of sequences,

all of which are derived from its predecessor in single steps or jumps. Such a walk may be either random or adaptive and will be guided by increasing desired mutant characteristics. An evolutionary trajectory may be considered as the path between the initial and final sequence. The length of the trajectory is usually equal to the number of single point mutations occurring in the evolutionary process.

From analysis of directed evolution experiments it has been found that the distance in evolutionary space that can be traversed in experiments is usually less than half a dozen steps. This can either be interpreted from the viewpoint that the sequences of the improved properties lie close to the original sequence; they are not sparsely scattered in sequence space. On the other hand it could indicate the ability of current methods to generate functional sequences more than a few point mutations away from the original sequence is limited (Miyazaki & Arnold, 1999).

Even if sequence space is to be sampled to a distance of only a few mutations, a strategy that relies on finding a successful variant by exhaustive searching of even a single sequence pool is likely to fail. The reason being a combinatorics problem; the codon dilemma. It is known amino acid sequence changes are encoded on a nucleotide level, thus a single residue change may require up to three co-ordinated changes of the coding sequences. It has been calculated that even a library of 10^8 sequences will not exhaustively encode all single point mutations. To densely sample sequence space to a distance of only one, or at absolute maximum, two, mutations the successful application of directed evolution requires that a path exists which will yield a considerably improved function with each small evolutionary step. This is analogous to natural

evolution; they both operate by arriving at the target sequence in stepwise improvements (Steipe, 1999).

To analyse an enzyme of 250 residues, if all the 20 natural amino acids were to be randomised completely into the chain an inordinately high number of permutations would result, 20^{250} mutant enzymes. This would require a mass greater than that of the entire universe (Reetz & Jaeger, 2000). Even if the entire mass of the earth was converted into starting material, only a 38-mer random protein could be created (Nixon & Firestine, 2000).

Mandecki published a paper entitled 'The game of chess and searches in protein sequence space' in 1998 (Mandecki, 1998). This paper focuses on the analogy between directed evolution and the game of chess and proposes that sequence space and chess space may be considered approximately equivalent. It is proposed that a complete and exhaustive search for the 'perfect' protein is equivalent to the search for the perfect chess game; both are equally unachievable either in the laboratory or in nature.

A considerable section of the paper considers the super-computer Deep Blue that defeated a world chess champion. The question is asked "if we can ourselves, and even develop a computer to, play chess at an advanced level, why can we not even play a beginners game in sequence space?" An explanation may be that we do not have the fitness functions available to evaluate random protein sequences, or sections of them, for attributes such as folding, structure and stability or even binding. Further we do not yet have algorithms which will reliably predict the three-dimensional structure, or even the function, of a protein purely from its sequence.

There are two different categories directed evolution techniques may be classified in. The first to be discussed here are the non-recombinant methods, which include error prone PCR and the exploitation of bacterial strains specifically prone to random mutagenesis. The second category encompasses the recombinant methods, including DNA and family shuffling, staggered extension process and random priming recombination. These shall be dealt with in turn.

Non-Recombinant Methods

Bacterial Mutator Strains

When wild-type *E. coli* DNA is propagated on a pBluescript derived plasmid after approximately 30 generations of growth it is found to exhibit a spontaneous mutation frequency of typically 2.5×10^{-4} per 1000 nucleotides.

A commercially available *E. coli* strain, XL1-Red, contains mutations in three independent DNA repair pathways and exhibits a spontaneous mutation frequency of 0.5 mutations per 1000 nucleotides. To generate a mutant library a gene encoding the enzyme to be manipulated can be cloned into an appropriate plasmid, transformed into the XL1-Red cells and grown overnight. The use of this strain is however somewhat limited; the absence of three DNA repair pathways result in it being relatively unstable. Thus if the target DNA is relatively small, less than 100bp, or if multiple mutations are required, the number of generations needed for propagation of the plasmid DNA becomes impossible due to cell degradation and ultimately death (Reetz & Jaeger, 1999).

Error Prone Polymerase Chain Reaction (epPCR)

For most purposes the introduction of a nucleotide change by epPCR would be the method of choice; it is simple, efficient, restricted to the area of interest, better characterised than some other directed evolution protocols and as such well documented in scientific literature (Cadwell & Joyce, 1994; Leung *et al.*, 1989). PCR was originally designed to amplify a given strand of DNA with extremely high fidelity; the intrinsic proof-reading characteristic of DNA polymerase ensures the amplification proceeds with almost 100% accuracy. In the course of a standard PCR reaction the wrong nucleotide may be incorporated during the amplification resulting in mutations occurring at a frequency of between approximately 0.1 and $2x10^{-4}$ per nucleotide when *Taq*-polymerase is used. It was found that this mutation frequency could be increased to a rate of $7x10^{-3}$ per nucleotide by several methods. Firstly the concentration of $MgCl_2$ can be increased from 1.5mmol L^{-1} to 7mmol L^{-1}. Similarly introducing 0.5mmol L^{-1} of $MnCl_2$ has also been shown to be effective and the concentrations of dCTP and dTTP may be increased from 0.2mmol L^{-1} to 1mmol L^{-1}. The amount of *Taq*-polymerase used is also increased from 2.5units to 5units to further promote the frequency of errors. (May *et al.*, 2000). Overall a combination of all of these methods is used in each experiment to induce the highest frequency of errors.

The conditions noted above may be manipulated to achieve the desired average mutation frequency of typically one or two per gene. This will result in an average exchange of one amino acid per mutant enzyme (Reetz & Jaeger, 1999).

Recombinative Methods

DNA Shuffling

The most successful approach to directed evolution to date may be considered to be DNA shuffling (Stemmer, 1994). This protocol was pioneered to address the question of how an efficient walk on an evolutionary landscape could be generated experimentally. The protocol allows successful mutations to be passed among sequences by recombination. The actual power of this process comes from the possibility of preserving locally optimal residues, for example the active site or catalytic residues (Altamirano *et al.*, 2000; Lebbink *et al.*, 2000; Merz *et al.*, 2000).

A simplified description of the methodology of DNA shuffling is given here, figure 1.08; more complete descriptions are available (Stemmer, 1994). One or more closely related genes are digested with DNAseI to yield double stranded oligonucleotide fragments, usually of between 10-50 base pairs (bp), which are then purified and used in a PCR style reaction. Repeated cycles of strand separation and reannealing, in the presence of DNA polymerase, and a final PCR amplification step results in the reassembly of a full-length gene. The actual recombination occurs via template switching, where fragments from one copy of the gene prime on another copy.

The combination of DNA shuffling with non-recombinative methods, such as epPCR, allows the simultaneous permutation of mutations from both single residues and large segments of DNA sequences. Once the mutants have been screened or selected for the desired property the mutant genes can be reshuffled, either against other mutant genes or the wild-type gene. This method corresponds to the backcross technique from classical genetics and is a process that may eventually lead to an elimination of neutral and deleterious mutations from the gene pool (Steipe, 1999).

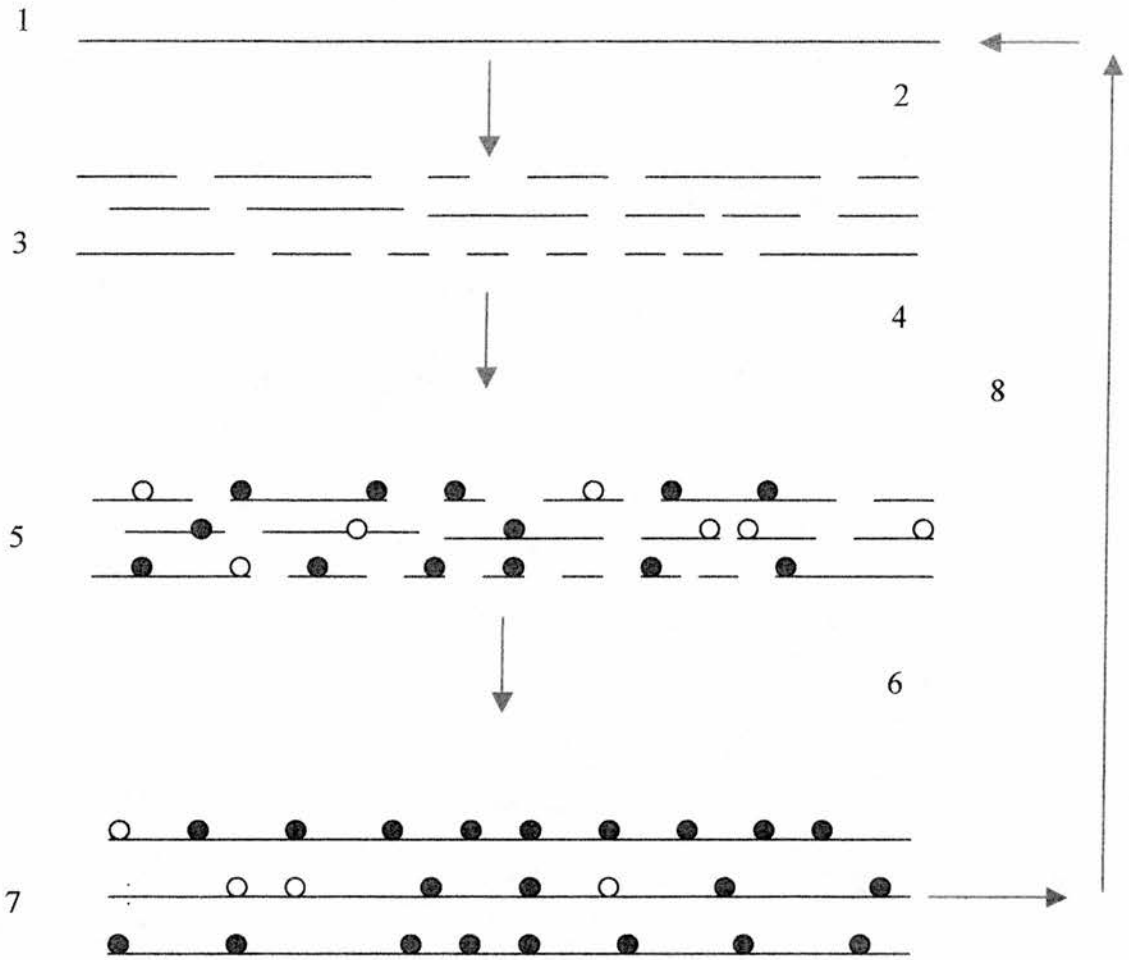


Figure 1.08. Diagrammatic representation of DNA shuffling method. The steps are;

1 Single gene
2 Random fragmentation
3 Pool of random fragments
4 Random DNA fragments
5 Reassembly
6 Selection & screening
7 Mutant genes with positive or negative clones
8 Cycle link

Family Shuffling

In this technique genes encoding enzymes which share the same fold but represent a different evolutionary pathway replace a single gene as the starting material. The technique is often regarded as 'sexual PCR;' the related genes are amplified under conditions favourable for crossover or intergenic recombination (Crameri *et al.*, 1998). The most important feature of this directed evolution technique is because the related enzymes share a common fold any chimeric polypeptides produced are likely to be active due to them being correctly folded. Family shuffling reaches a far greater portion of sequence space than is possible with single gene shuffling and thus accelerates the rate of functional enzyme improvement. It is currently considered to be at the very forefront of combinatorial technology and directed evolution research. This method is now replacing epPCR as the procedure of choice for researchers at the cutting edge of this technology (Kikuchi *et al.*, 2000).

Staggered Extension Process (StEP)

This process is similarly based on a staggered PCR-like reaction and is a simple and efficient method for the *in vitro* mutagenesis and recombination of targeted genes (Zhao *et al.*, 1998). One defined primer is added to the template DNA, which comprises of two or more genes. An overtly abbreviated primer extension reaction, catalysed by DNA polymerase, produces short DNA fragments which can anneal to different templates, these are further extended during the next short cycle of primer extension. This process is repeated until full-length genes are formed which can be amplified in a conventional PCR experiment using external primers.

Random Priming Recombination (RPR)

In RPR random primers are used to generate short DNA fragments, 50-500bp, which are complementary to different segments of a target DNA (Shao *et al.*, 1998). In the next step these fragments, which also carry mutations as a result of mispriming and base misincorporation, are allowed to prime each other. This results in recombination during reassembly to full-length genes, which is carried out by repeated thermocycling in the presence of thermostable DNA polymerase. It is the opinion of some that this method has advantages over DNA shuffling because it uses single stranded templates and thus, in theory, every nucleotide of the template should be mutated at an equivalent frequency.

Screening & Selection

Screening is the process of identifying a desired member of a library in the presence of all the other members. Selection is the enrichment of active variants in a molecular library (Reetz & Jaeger, 1999). It is known for libraries of 10^{10} mutant genes to be produced by directed evolution techniques with relative ease (Steipe, 1999). Therefore the development and application of an appropriate search system is of exceptional importance. Indeed directed evolution as a whole may be thought to revolve around, and thus be inherently dependent upon, the process of the selection and screening of generated mutant genes.

After screening, if selection is applied, only the desired member of a library appears, for example, as a viable microbial clone. This '*in vivo* selection' is usually very efficient because it allows only micro-organisms to grow which express a gene encoding a particular enzyme essential for survival. Unfortunately it can be difficult to develop

such a system because microbial cells are extremely versatile with respect to circumventing environmental restrictions.

High-Throughput Screening (HTS)

The first high-throughput screening system for enantioselectivity was developed specifically for reactions involving lipase-catalysed kinetic resolution of chiral esters. In this system *p*-nitrophenol esters are used; hydrolysis produces *p*-nitrophenol which can be detected by UV/Vis spectroscopy (Reetz & Jaeger, 2000). Between 500 and 600 mutants can be screened by one person each day using this method. Fluorescence can also be applied to this methodology, providing the substrate has fluorogenic functionality. The additional benefit of fluorescence screening is that only small amounts of substrate are required for testing.

Potentially the most broadly applicable and exact method for HTS of enantioselectivity involves the use of isotopically labelled substrates in the form of pseudo-enantiomers or pseudo-prochiral compounds (Jaeger & Reetz, 2000). The first family, the pseudo-enantiomers, are chiral compounds that differ from one another in that they have opposite absolute configuration and isotopic labelling. The second family, the pseudo-chiral compounds, are derived from their corresponding prochiral compounds and contain isotopic labelling in one of the enantiomeric groups. The screening method utilises a detection system, such as mass spectrometry (MS) combined with electrospray ionization (EI). This method enables the measurement of the relative amounts of isotopically labelled pseudo-enantiomers, which correspond to true enantiomers, by integrating the MS peaks. It has been reported that approximately 1000 exact

enantiomeric excess (*ee*) values can be determined by one person each day using this method (Reetz *et al.*, 1999).

Alternative Detection Methods

Novel strategies include the measurement of black-body radiation as a function of enantioselectivity. For example, the lipase-catalysed hydrolysis of *R* and *S* configured acetates can be screened with the aid of an IR thermographic camera; hot spots indicate exothermic reactions on the microtitre plate. Differences between the two forms can be easily detected (Reetz *et al.*, 1998). However, a more sensitive method remains to be developed to enable small improvements in enantioselectivity to be detected. Circular dichroism has the possibility of being used in selection and screening although to date it has not been applied to enzymatic reactions (Ding *et al.*, 1999).

Concluding Remarks

Aldolases are important reagents for the organic chemistry by virtue of their stereo control of bond formation. Directed evolution is a powerful method of adapting enzymes to carry out a specific function. This thesis describes the structural characterization of KDPG aldolase from *E. coli*, and the structure of a novel aldolase made by directed evolution.

Chapter Two

Over-expression, purification and crystallisation of KDPG Aldolase from Escherichia coli

2.1 Abstract

2-keto-3-deoxy-6-phosphogalactonate aldolase was initially received as partially purified frozen glycerol stocks. The sample was further purified by HQ chromatography giving a single band on analysis by silver stained SDS-PAGE. Dynamic light scattering showed the protein to be mono-disperse and molecular weight analysis indicated it to exist as a trimer in solution. Crystals were obtained under several Hampton screen I and II conditions.

To improve protein purity the plasmid containing KDPG aldolase in a pBluescript vector was transformed into the BL21(DE3) cell line. A two-step HPLC purification method was used to purify the sample to a single band on silver stained SDS-PAGE and IEF gels. Optimisation of crystallisation conditions gave crystals which diffracted to 2.8Å in-house. Exhaustive attempts to find a molecular replacement solution using a 2.2Å P. putida model were unsuccessful. These co-ordinates were provided by Professor Tulinsky.

A 2.17Å data set was collected at the CLRC Daresbury synchrotron on station PX9.6. Further exhaustive attempts to find a solution in AMoRe using this new high-resolution data were unsuccessful. Attempts to find a solution using another molecular replacement program, EPMR, were also unsuccessful.

2.2 Introduction

As mentioned the Entner-Doudoroff pathway consists of two enzymes; 6-phosphogluconate dehydratase encoded by the *edd* gene, and KDPG aldolase encoded by the *eda* gene. The pathway in *E. coli* is used for gluconate metabolism.

Whilst extensive studies on the molecular biology of this pathway have been carried out in *Z. mobilis*, very few investigations have focused on the *E. coli* pathway. In *E. coli* the *edd* and *eda* genes are linked to the *zwf* gene which encodes for glucose-6-phosphate dehydrogenase. These genes are tightly regulated and are controlled completely independently. Expression of the *eda* gene was found to be subject to control by the *kdg*R product. (Egan *et al.*, 1992).

Selection of the plasmid-borne functional *eda* gene was accomplished by complementation of defective gluconate metabolism in *E. coli* RW231 on minimal media. The plasmid, pLC37-44, was transferred into *E. coli* RW231 by conjugal mating. Constructions containing the intact *eda* gene in opposite orientations, designated pTC190 and pTC 191 respectively, were made by subcloning a 1.5kb *NruI* restriction fragment from pLC37-44 into the *Eco*RV site of pBluescriptII. The *eda* gene reading frame was found to be 642 bases long and encode the 213 residue, 22 283Da KDPG aldolase enzyme (Egan *et al.*, 1992). Alignment of the *E. coli* enzyme with those from *Z. mobilis* and *P. putida* shows it to be 52% and 44% identical respectively, with a similarity of 68% and 65% respectively. The sequence alignment of the *P. putida* and *E. coli* enzymes is shown in figure 2.01.

The structure of KDPG aldolase from *P. putida,* shown in figure 2.02, was solved to 3.5Å in 1976 (Mavridis & Tulinsky, 1976) and refined to 2.8Å in 1982 (Mavridis *et al.*, 1982). Studies on the synthetic usefulness of the enzyme in organic chemistry prompted further studies of its structure. Initially the enzymes isolated from *P. putida*, *Z. mobilis* and *E. coli* were used to determine which was most suitable for use with non-natural substrates in conditions unlike those in the cell. KDPG aldolase from *E. coli* was chosen (Shelton *et al.*, 1996). The results they observed from experiments did not agree with the structural observations made from the *P. putida* 2.8Å structure determined in 1982.

It was decided to determine the structure of KDPG aldolase from *E. coli* with a view to being able to engineer non-natural substrate specificity into the active site. It was thought the structure solution would be straightforward; the *P. putida* data was determined to 2.8Å and has almost 45% sequence identity with the *E. coli* enzyme. The approach to be taken was to use the molecular replacement program *AMoRe*, searching for a solution of the *E. coli* enzyme using the *P. putida* data as the search model. Although only the 3.5Å carbon-α co-ordinates are deposited in the protein data bank, Professor Tulinsky provided us with 2.2Å refined co-ordinates.

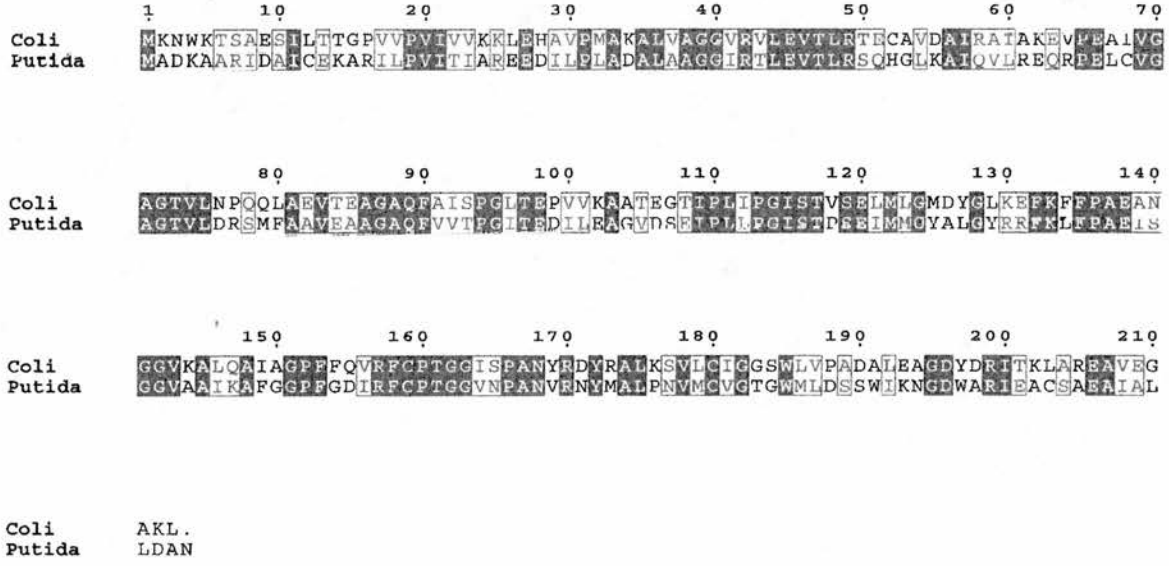


Figure 2.01. Sequence alignment of KPDG aldolase from *E. coli* and *P. putida*. Residues in red are identical; residues in blue boxes are similar.

Figure 2.02. *P. putida* KDPG aldolase determined to 3.5Å in 1976 and deposited in the protein data bank. As may be clearly seen this structure has minimal resemblance to a TIM barrel.

2.3 Experimental

Initial Samples

The first samples of KDPG aldolase obtained were sent as frozen glycerol stocks on dry ice from our collaborators, the group of Professor E.J. Toone, Duke University, USA. The samples were defrosted on ice and dialysed extensively into 20mM Tris-base, pH 8.0, to remove the glycerol which past experience has shown to be problematic during HPLC purification. The KDPG aldolase was of sufficient purity (figure 2.03) for its usage as an enzyme catalyst in synthetic organic chemistry reactions, however it was not sufficiently pure for crystallisation trials to commence.

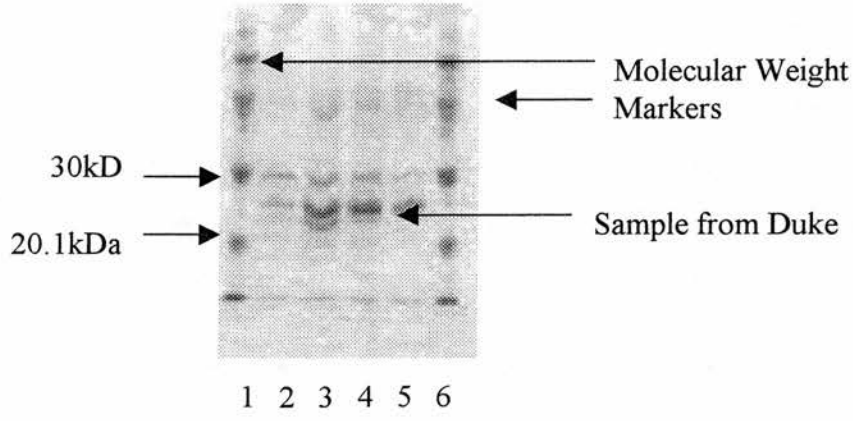


Figure 2.03. SDS PAGE gel of KDPG aldolase received from Duke University. Lanes 1 & 6 are molecular weight markers. Lanes 2-5 are samples from Duke at varying protein concentrations.

The dialysed sample was 0.2μm filtered and applied to a POROS-HQ anion exchange HPLC column, BioCad Sprint, equilibrated in 20mM Tris-base pH 8.0. The protein was eluted with a 0-1000mM NaCl, 20mM Tris-base pH 8.0, salt gradient. All fractions were collected and analysed by SDS-PAGE; Coomassie staining showed one fraction

with a single band at approximately 23kDa, the theoretical molecular weight of KDPG aldolase (figure 2.04). The fraction was concentrated under nitrogen pressure in an Amicon ultrafiltration unit. After dialysing into 20mM Tris-Base pH 8.0, a silver stained IEF gel was run on the sample to analyse charge homogeneity; this showed a single band.

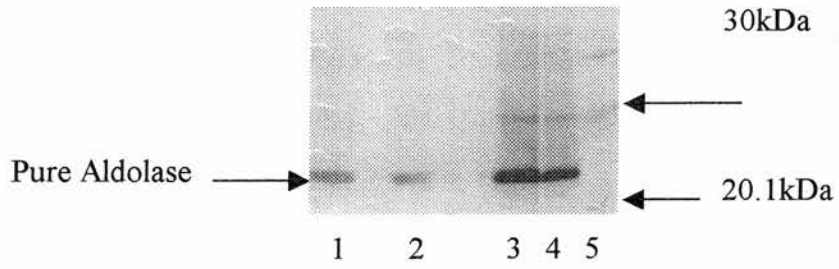


Figure 2.04. SDS PAGE gel of pure KDPG aldolase. Lanes 1 & 2 show pure KDPG aldolase. Lanes 3&4 show heavily over-loaded KDPG aldolase prior to further purification. Lane 5 is the molecular weight markers.

Dynamic light scattering (DLS) was performed on the KDPG aldolase at a concentration of 5mg ml^{-1}, as determined by the Bradford method (Bradford, 1976). This experiment showed the protein to be mono-disperse, shown by a base line of 1.000, and thus likely to crystallise. It also showed the estimated molecular weight to be between 52 and 56 kDa suggests the protein exists as a trimer in solution. A dimer would give 44kDa as its estimated weight, a trimer 66kDa. The results of the DLS are shown in table 2.01.

Amplitude	Differential Coefficient (nm)	Radius (nm)	Polydispersity (nm)	Estimated Weight (kDa)	Count Rate	Base Line	SOS Error
0.850	589	3.2	/	52	105	0.999	0.47
0.852	586	3.3	0.542	53	104	1.000	0.43
0.853	588	3.2	/	52	103	1.001	0.31
0.836	586	3.3	0.548	54	103	1.002	0.38
0.849	585	3.3	/	54	101	1.000	0.43
0.830	577	3.3	0.767	56	104	1.004	1.07
0.841	591	3.3	/	54	102	1.001	0.25

Table 2.01. DLS results on pure KDPG aldolase

Initial crystallisation screens were set up using the following screens; Hampton I & II, MPD, PEG 6K and PEG 6K/LiCl, using KDPG aldolase at 16mg ml^{-1}. The trials were incubated at 293.5K. Analysis after 48 hours showed a number of conditions to give small needle shaped crystals. The conditions found are shown in table 2.02.

Screen			
Hampton I	0.2M Ammonium Acetate	0.1M Sodium Acetate pH 4.6	30% PEG 4K
	0.2M Magnesium Acetate Tetrahydrate	0.1M Cacodylate pH 6.5	30% MPD
	0.2M Ammonium Sulfate	-	30% PEG 8K
	0.2M Ammonium Sulfate	-	30% PEG 4K
Hampton II	0.5M Ammonium Sulfate	0.1M Sodium Citrate pH 5.6	1M Lithium Sulfate
PEG 6K	10% PEG	0.1M Citric Acid pH 4	-
	20% PEG	0.1M Citric Acid pH 5	-
	30% PEG	0.1M Mes pH6	-
	30% PEG	0.1M Hepes pH 7	-
	30% PEG	0.1M Tris pH 8	-
PEG 6K/LiCl	20% PEG	1M Lithium Chloride pH 4	-
	30% PEG	1M Lithium Chloride pH 4	0.1M Citric Acid

Table 2.02. Initial crystallisation screen hits for KDPG aldolase

Given this promising result it was decided the protein should be expressed in St Andrews; this would allow optimisation of the yield and purity of the protein for crystallisation.

Plasmid Transformation

KDPG aldolase plasmid DNA (pTC191) was transported on dry ice from the collaborators at Duke University to St Andrews. The plasmid was transformed into BL21(DE3)pLysS cells. The plasmid at this point was in a Bluescript II vector which is a non-inducible vector typically used in screening programmes to isolate DNA.

The transformed BL21(DE3)pLysS cells were grown on an Luria agar plate containing 100μg ml^{-1} ampicillin and 50μg ml^{-1} chloramphenicol, from this a single colony was picked and used to inoculate an overnight culture of 10ml Luria Broth (LB) plus 100μg ml^{-1} ampicillin and 50μg ml^{-1} chloramphenicol, incubated at 310K, 175rpm shake. This culture was then used to inoculate 500ml of LB plus 100μg ml^{-1} ampicillin and 50μg ml^{-1} chloramphenicol and incubated at 310K, 175rpm shake. When using a non-inducible plasmid the cells are typically harvested in the middle of the *E. coli* growth logarithmic phase to ensure the maximum level of protein expression. To determine this stage 1ml samples were taken every 15 mins for 4 hours and centrifuged at 13000 g for 3 mins. The resultant pellet was resuspended and analysed by Coomassie stained SDS-PAGE. This showed the optimum incubation time to be 3 hours.

Following these preliminary studies 2L of KDPG aldolase was prepared in an identical manner. The resultant cells were harvested by centrifugation; 8000 g, 15 mins, 277K, and the pellets frozen at 253K. Immediately prior to purification the pellets were defrosted on ice and resuspended in ice-cold suspension buffer, 20mM Tris-base pH 8.0, 1mM EDTA and 1mM PMSF, at a concentration of 10ml L^{-1} cell culture. The solution was sonicated on ice, 6 cycles of 30 s with 30 s rest period. The cell debris was

removed by centrifugation; 18000 g, 15 mins, 277K and the resulting solution 0.2μm filtered.

Crude KDPG aldolase was applied to a POROS-HQ HPLC column, BioCad Sprint, equilibrated in 20mM Tris-base pH 8.0. The protein was eluted with a 0 to 1000mM NaCl, 20mM Tris-base pH 8.0, salt gradient. The previously identified fractions anticipated to contain the KDPG aldolase were eluted at 125mM on the salt gradient, however all fractions were collected and analysed. A typical profile trace is shown in figure 2.05. Coomassie stained SDS-PAGE showed the fraction eluted at 125mM salt to contain a protein of 23kDa which was identified as being KDPG aldolase by comparison with the pure protein initially used in the crystallographic screening.

The fractions eluted at 125mM salt were pooled and the protein was dialysed exhaustively into 20mM Na_2HPO_4 pH 8.0 to remove any NaCl present. The sample was then concentrated under nitrogen pressure in an Amicon ultrafiltration unit. $(NH_4)_2SO_4$ was gradually added to 40% saturation, whilst stirring at 277 K. The solution was 0.2μm filtered and applied to a POROS high density phenyl HPLC column, BioCad Sprint, equilibrated in 40% $(NH_4)_2SO_4$, 20mM Na_2HPO_4, pH 8.0. The protein was eluted from a decreasing salt gradient of 40 to 0 % $(NH_4)_2SO_4$, 20mM Na_2HPO_4, pH 8.0, the fractions were collected and analysed by silver stained SDS-PAGE, which gave a single band. A typical profile trace may be seen in figure 2.06. Again the fractions containing KDPG aldolase were pooled, dialysed extensively into 20mM Tris-base, pH 8.0, and concentrated under nitrogen pressure and in an Amicon ultrafiltration unit.

At this stage the initial crystallisation screens were repeated; 1μl protein plus 1μl mother liquor, 100 μl reservoir exploiting the sitting-drop diffusion method; crystals were again obtained under several conditions. To ensure it was KDPG aldolase crystals a selection of them were crushed and dissolved in water to allow them to be N-terminal sequenced. The results showed the crystals to be KDPG aldolase and confirmed the N-terminal was fully intact.

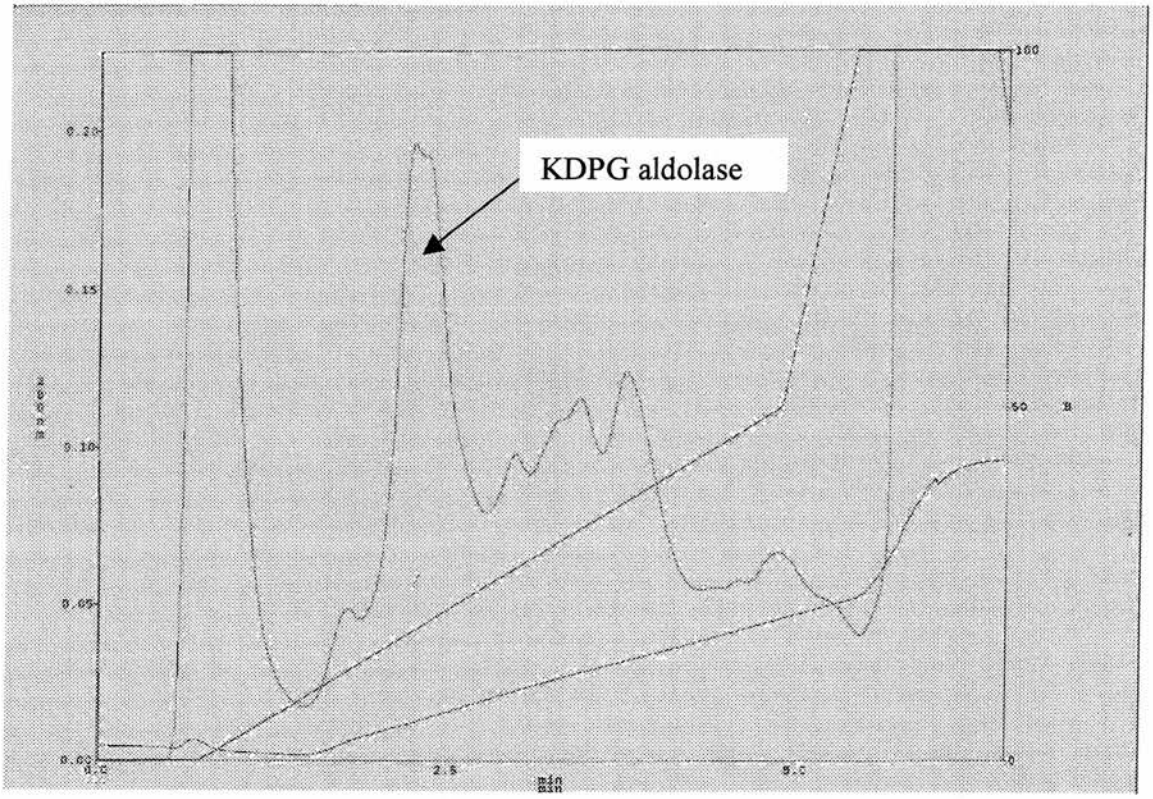


Figure 2.05. UV_{280} trace of the protein fractions eluted during anion exchange chromatography. Time is on the x-axis, left y-axis is 280nm reading, right y-axis is % of NaCl. The peak at 2.5 mins on the gradient is the target protein.

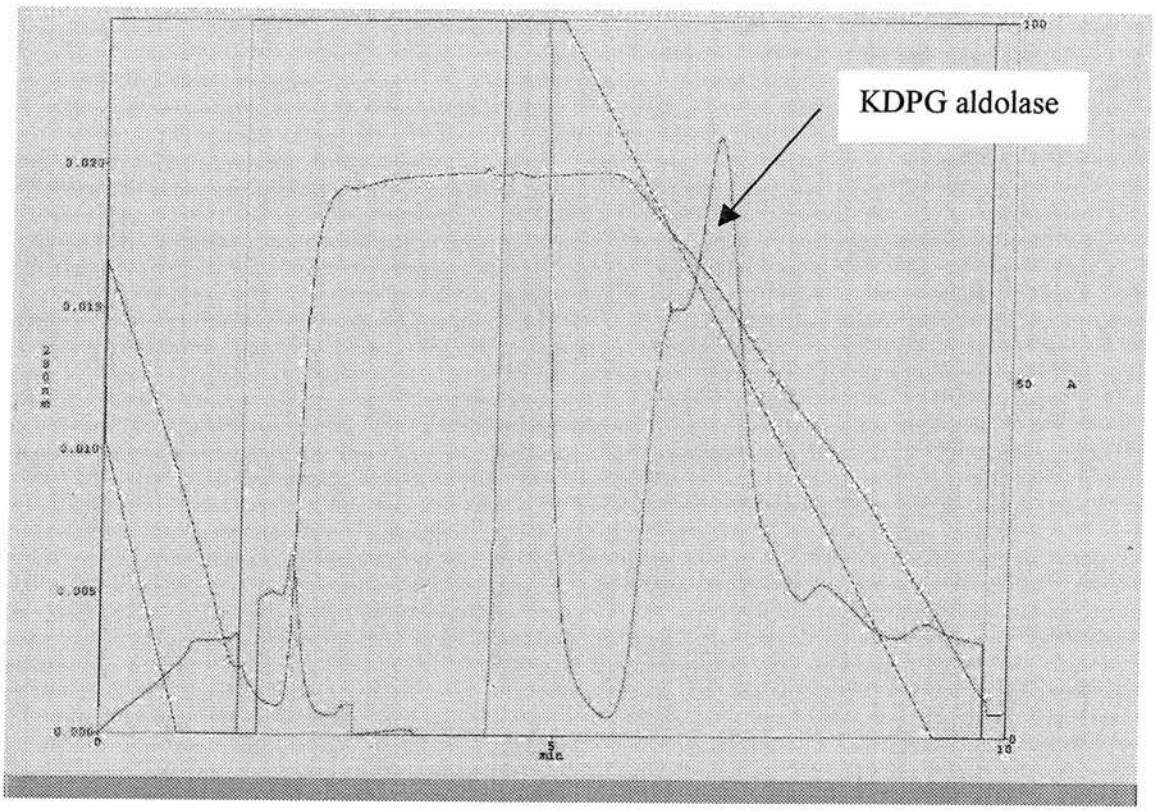


Figure 2.06. UV_{280} trace of the protein fractions eluted during hydrophobic exchange chromatography. Time is on the x-axis, left y-axis is 280nm reading, right y-axis is % of $(NH_4)_2SO_4$. The peak on the middle of the gradient is the target protein.

Optimisation of the crystallisation conditions followed. Conditions initially optimised included protein concentration, precipitant concentration, buffer pH and the method of crystallisation used. The conditions which gave the best crystals were found to be 20% PEG 6K, 0.075M citric acid pH 4.0, with a protein concentration of 3mg ml^{-1} as determined by Bradford (Bradford, 1976), using the sitting drop vapour diffusion method with 2µl protein plus 2µl mother liquor drop size. The crystals grew overnight at 293.5K. In an attempt to slow the crystal growth, with the aim of obtaining larger crystals, plates were laid out and kept at 277K. However, this was found not to have any

difference on either the crystal size nor the time taken to grow. Figure 2.07 shows the crystals initially obtained from the Hampton screening.

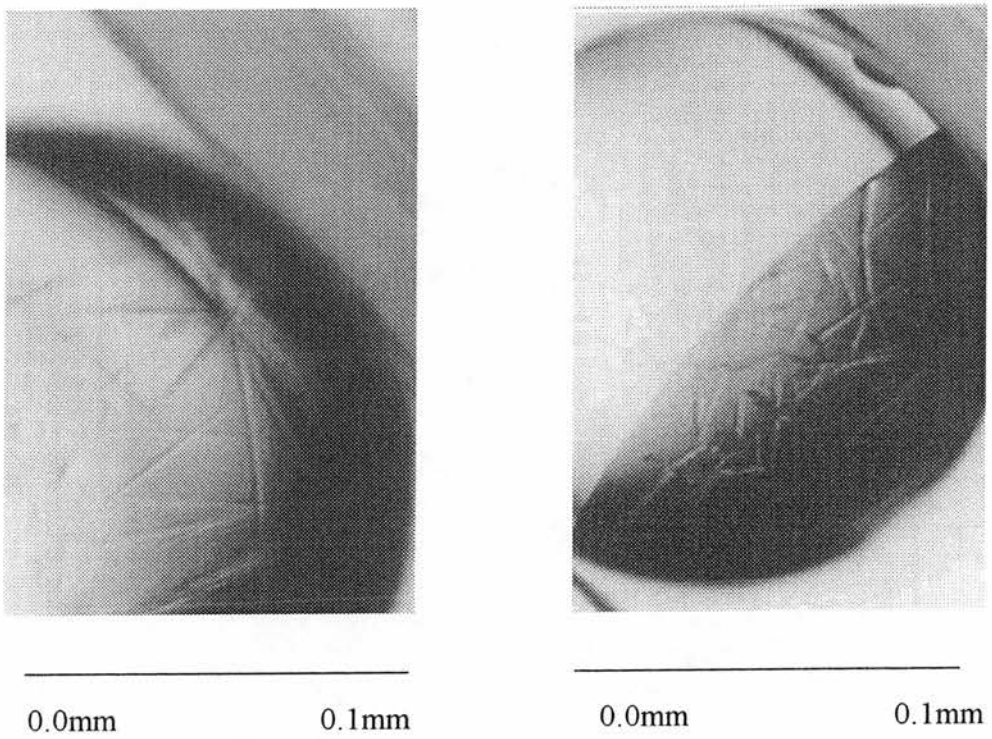


Figure 2.07. Initial crystals obtained from Hampton I & II

Data Collection

Copper Rotating Anode

Initial data were collected from a single crystal in-house using a Nonius DIP2000 dual image plate. The crystal was soaked for 2 mins in the mother liquor plus 15% glycerol prior to data collection at 130K. Data were collected as 90 images with an exposure time of 30 mins and a 1° non-overlapping oscillation. The crystal to detector distance was 200mm. Figure 2.08 shows a typical KDPG aldolase diffraction pattern. Data were indexed and integrated with *DENZO*, version 1.12d, and merged with *SCALEPACK* (Otwinowski, 1996) to a resolution of 2.8Å. The data were 80% complete, gave an R_{merge} of 7.5% with a 37% solvent content. The cell constants were determined as a = 56.3Å, b = 95.3Å, c = 136.2Å, $\alpha = \beta = \gamma = 90°$. Indexing and merging identified the crystal as a member of the Laue class, the possible space groups being $P222_1$, $P2_12_12_1$ or $P2_12_12$. As only one axis with an n=2 condition was clearly observed, it was not possible to be more definite regarding the identity of the space group.

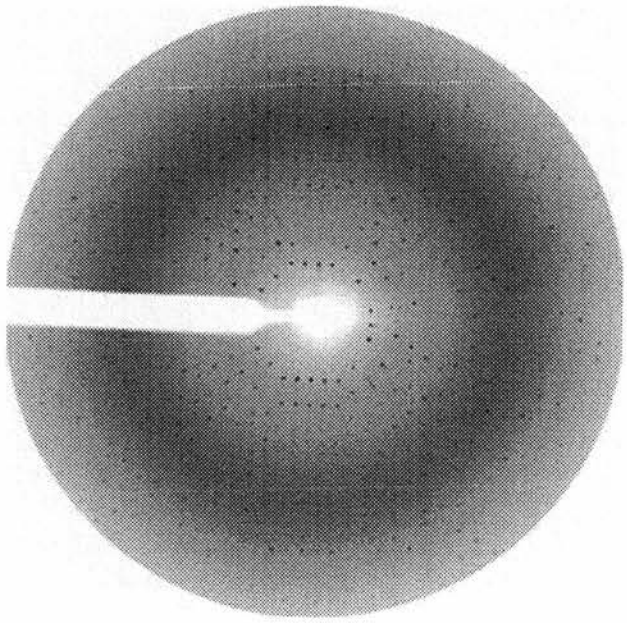


Figure 2.08. Native KDPG aldolase in-house diffraction pattern

Synchrotron Radiation

Data were collected from a single crystal on station PX9.6 at the CCLRC synchrotron at Daresbury using an ADSC Quantum-4 CCD detector. The crystal was soaked for 2 mins in the mother liquor plus 15% glycerol prior to data collection at 100K. Data were collected as 180 images with an exposure time of 90 s and a 1° non-overlapping oscillation. The crystal to detector distance was 230 mm using a fixed wavelength of 0.87Å.

Data were indexed and integrated with *DENZO*, version 1.12d, and merged with *SCALEPACK* (Otwinowski, 1996) to a resolution of 2.17Å. However, completeness at this resolution was affected by the square geometry of the detector. Data were indexed in the orthorhombic space group $P2_12_12_1$ following observations of a *2n* condition along all three axes; at least 11 systematic absences were measured for each axis. The data had a mosaic spread of 0.8° and gave an R_{merge} of 5.7% at 2.26Å resolution, with unit cell parameters a = 53.2Å, b = 77.79Å & c = 146.83Å, $\alpha = \beta = \gamma = 90°$. A dimer in the asymmetric unit gives a solvent content of 62%; a trimer in the asymmetric unit gives a solvent content of 44%. A complete summary of the data is given in table 2.03.

Resolution (Å)	No. Reflections	Completeness (%)	R_{merge} (%)	Redundancy
30.0 – 4.52	3899	100	4.3	6.4
4.52 – 3.59	3698	99.4	4.6	6.0
3.59 – 3.14	3528	97.0	5.7	5.4
3.14 – 2.85	3535	96.2	7.2	5.4
2.85 – 2.65	3439	95.1	9.2	5.3
2.65 – 2.49	3361	93.0	10.3	4.9
2.49 – 2.37	3209	88.5	12.3	4.0
2.37 – 2.26	2932	81.3	13.1	2.8
2.26 - 2.18	2205	61.5	14.4	3.0
2.18 – 2.10	1380	38.7	16.4	1.4
30.0 – 2.10	31186	85.3	5.8	4.7

Table 2.03. 2.1Å KDPG aldolase data statistics

Molecular Replacement

It was planned to determine the structure of KDPG aldolase by molecular replacement as implemented in the CCP4 program AMoRe (CCP4, 1994; Navaza & Saludjian, 1997). At this stage of the investigation it was not completely clear if KDPG aldolase existed in a dimeric or trimeric form. The unit cell of the crystal gave a Matthews number of 3.3Å^3 Da^{-1} for a dimer in the asymmetric unit and 2.2Å^3 Da^{-1} for a trimer; both of which are feasible. Self-rotation analysis was also unclear because the space group has no two-fold axis; one-and-a-half dimers in the asymmetric unit are not feasible.

KDPG aldolase from *E. coli* shares 45% sequence identity with KDPG aldolase from *P. putida*; it is implausible the *E. coli* enzyme does not share the same fold as the trimeric *P. putida* enzyme.

The KDPG aldolase carbon α co-ordinates from *P. putida* were deposited to only 3.5Å in the PDB; the structure was later solved to 2.8Å and whilst this structure was published the co-ordinates were not deposited. Professor Tulinsky very kindly provided us with 2.2Å refined, but not deposited or published, complete co-ordinates of this enzyme.

Initial efforts to find a molecular replacement solution using the trimer found in *P. putida* and the 2.8Å *E. coli* data as a search model were unsuccessful, even when variable loops were excluded and non-conserved side chains were removed. Extensive testing of resolution ranges also failed to provide a solution using a trimeric search model. The next approach was to use a monomer from the *P. putida* as a search model; this gave no solutions. Initially it was thought the low resolution and completeness of the data collected in-house may have caused some problems thus the method was repeated with the 2.17Å synchrotron data when it became available. However, a solution remained elusive. It is not possible to list the details of all the searches which were performed; tables 2.04 to 2.11 give the 10 best solutions from the rotational search and the equivalent solutions from the translational search for a selection of search strategies. The values shown in these tables are typical of those obtained for all the searches performed.

Results obtained using 2.8Å *E. coli* data against 2.8Å *P. putida* trimeric poly-alanine model using the resolution range 15 to 3Å in AMoRe are shown in tables 2.04 and 2.05. Throughout the correlation coefficient is the correlation coefficient between the observed amplitudes for the crystal and the calculated amplitudes for the model. The R_{factor} is the classic R_{factor} between the observed intensities for the crystal and the sum of the calculated intensities for all the symmetry equivalents of the model. α, β, γ, are the Eulerian angles and x, y, z, define the molecules position in the cell.

Solution	α	β	γ	c.c*
1	167.47	57.35	311.91	9.0
2	167.69	57.44	71.13	8.8
3	167.84	57.6	191.54	8.7
4	33.87	77.01	205.51	8.6
5	13.38	65.75	126.84	8.4
6	13.38	65.53	246.92	8.4
7	33.81	77.19	85.54	8.4
8	33.60	76.99	325.76	8.2
9	13.32	65.63	6.86	8.1
10	97.74	67.39	249.59	7.7

* c.c is the correlation coefficient.

Table 2.04. Solution from rotation function searching

Solution	α	β	γ	x	y	z	c.c*	R_{factor}
1	168.67	87.83	189.73	0.4823	0.3678	0.4754	4.0	58.1
2	168.67	87.83	189.73	0.4807	0.4295	0.2317	3.7	57.3
3	168.67	87.83	189.73	0.2524	0.3691	0.4758	3.5	58.2
4	168.67	87.83	189.73	0.4858	0.3671	0.2283	3.3	58.2

Table 2.05. Solution from translation function searching

Results obtained using 2.1Å *E. coli* data against 2.2Å *P. putida* trimeric poly-alanine model using the resolution range 12 to 4Å in AMoRe are shown in tables 2.06 and 2.07.

Solution	α	β	γ	c.c*
1	98.51	67.42	248.55	8.0
2	170.37	55.76	310.18	8.0
3	98.50	67.23	128.72	7.8
4	170.42	55.90	190.13	7.8
5	79.8	64.49	124.78	7.8
6	170.53	55.81	70.13	7.8
7	98.58	67.07	8.57	7.8
8	79.49	64.46	5.02	7.6
9	80.03	64.45	244.5	7.5
10	84.00	26.29	268.5	7.0

*c.c is the correlation coefficient

Table 2.06. Solutions from rotation function searching

Solution	x	y	z	c.c*	R_{factor}
1	99.81	66.67	248.11	7.4	58.4
2	99.64	66.65	128.21	7.4	58.4
3	99.87	66.54	8.11	7.4	56.3
4	170.47	56.27	310.13	15.6	58.3
5	80.16	64.55	124.12	7.5	56.3
6	170.5	56.47	190.12	15.5	56.3
7	170.67	56.34	70.08	15.4	56.3
8	79.84	64.52	4.3	7.7	58.3
9	80.41	64.53	243.8	7.6	58.1
10	35.45	76.14	82.5	8.1	58.7

Table 2.07. Solutions from translation function searching

A selection of solutions from the translation refinement were used to make a trimeric model of the protein. This was then subjected to CNS simulating annealing; the starting and final R_{factor} was 63.41% and 63.88%, the R_{free} values were 59.41% and 58.80%. These potential solutions would not refine and thus were discarded.

Results obtained using 2.1Å *E. coli* data against 2.2Å *P. putida* monomeric poly-alanine model using the resolution range 15 to 3Å in AMoRe are shown in table 2.08 and 2.09.

Solution	α	β	γ	c.c*
1	34.91	76.16	202.84	10.7
2	34.95	76.12	322.84	10.6
3	34.98	76.18	82.84	10.6
4	98.82	67.06	248.41	9.5
5	98.70	67.01	128.46	9.2
6	98.84	67.01	8.42	9.2
7	81.31	64.39	123.41	9.2
8	89.42	79.61	17.40	9.1
9	81.38	64.37	243.32	9.1
10	89.94	79.60	257.18	9.0

*c.c is the correlation coefficient

Table 2.08. Solutions rotation function searching

Solution	α	β	γ	x	y	z	c.c*	R_{factor}
1	34.91	76.16	202.84	0.0579	0.0591	0.0710	5.2	58.0
2	34.95	76.12	322.84	0.0567	0.0595	0.0074	5.2	58.0
3	34.98	76.18	82.84	0.0578	0.0953	0.0700	5.3	57.9
4	98.82	67.06	248.41	0.2480	0.4666	0.1240	4.2	57.4
5	98.70	67.01	128.46	0.2476	0.4668	0.1241	4.3	57.4
6	98.84	67.01	8.42	0.2472	0.4667	0.1240	4.3	57.4
7	81.83	64.39	123.41	0.0351	0.1437	0.4141	4.5	58.0
8	89.42	79.61	17.4	0.3887	0.3991	0.2772	4.8	57.5
9	81.38	64.37	243.32	0.0351	0.1441	0.4141	4.4	57.9
10	89.94	79.60	257.18	0.3883	0.3988	0.2773	4.8	57.5

Table 2.09. Solutions from translation function searches

Results obtained using 2.1Å *E. coli* data against 2.2Å *P. putida* monomeric poly-alanine model using the resolution range 12 to 4Å in AMoRe are shown in tables 2.10 and 2.11.

Solution	α	β	γ	c.c*
1	167.46	57.35	311.91	9.0
2	167.69	57.44	71.73	8.8
3	167.84	57.60	191.54	8.7
4	33.87	77.01	205.51	8.6
5	13.38	65.75	126.84	8.4
6	13.38	65.53	246.92	8.4
7	33.81	77.19	85.54	8.4
8	33.60	76.99	325.76	8.2
9	13.32	65.63	6.86	8.1
10	97.74	67.39	249.59	7.7

*c.c is the correlation coefficient

Table 2.10. Solutions from rotation function searches

Solution	α	β	γ	x	y	z	c.c*	R_{factor}
1	167.46	57.35	311.91	0.4815	0.1833	0.1265	6.9	56.8
2	167.69	57.44	71.73	0.4150	0.1840	0.1265	7.1	56.8
3	167.84	57.60	191.54	0.4170	0.1851	0.1264	7.1	56.7
4	33.87	77.01	205.51	0.0433	0.2979	0.1004	5.7	57.6
5	13.38	65.75	126.84	0.4818	0.2491	0.2999	5.4	56.9
6	13.38	65.53	246.92	0.4818	0.2494	0.2998	5.5	56.9
7	33.81	77.19	85.54	0.0431	0.2980	0.1003	5.9	57.6
8	33.60	76.99	325.76	0.3769	0.4872	0.3160	5.1	57.4
9	13.32	65.63	6.86	0.4819	0.2492	0.2997	5.4	57.0
10	97.74	67.39	249.59	0.2462	0.4685	0.1252	4.3	57.4

Table 2.11. Solutions from translation function searches

At this stage another molecular replacement package was used, EPMR. This is a program for crystallographic molecular replacement searches using an evolutionary algorithm (Kissinger *et al.*, 1999). Similar procedures were followed using this program as were applied in AMoRe. Still no solutions were found using monomeric and trimeric search models with the highest resolution data sets for *E. coli* and *P. putida.*

Solution	α	β	γ	x	y	z	c.c*	R_{factor}
1	66.3	129.20	198.64	13.01	28.47	120.15	13.8	64.4
2	318.03	35.68	306.89	9.92	52.99	80.62	13.1	65.2
3	341.10	57.84	239.47	42.30	5.62	45.05	13.1	65.2
4	325.93	94.98	8.41	19.47	13.37	63.73	13.1	65.2
5	206.54	85.53	339.41	0.31	9.12	76.28	13.0	65.6
6	331.32	83.19	298.49	31.51	62.32	63.39	12.9	65.2
7	344.11	117.94	126.79	5.21	28.92	47.57	12.8	64.8
8	254.63	159.60	121.86	31.95	6.55	47.18	12.8	64.9
9	176.71	38.95	344.65	16.24	55.94	47.66	12.8	64.9
10	72.26	52.67	35.06	16.40	55.84	80.47	12.7	64.8

*c.c is the correlation coefficient

Table 2.12. Solutions from monomer searching from 20 to 5Å with 2.1Å *E. coli* and 2.2Å *P. putida*

Solution	α	β	γ	x	y	z	c.c*	R_{factor}
1	247.87	113.51	322.67	-16.26	12.97	124.74	27.7	58.9
2	249.60	114.95	316.97	-123.67	12.90	124.45	27.7	58.9
3	254.08	114.73	322.81	36.99	12.80	124.20	27.7	58.9
4	210.25	29.06	182.89	202.45	12.90	22.63	27.7	58.9
5	220.65	31.98	172.02	202.44	12.99	22.03	27.7	58.9
6	216.67	31.79	181.69	43.75	12.88	22.43	27.7	58.9
7	247.85	113.75	142.11	42.66	25.82	51.58	27.4	59.0
8	249.99	114.57	138.50	-10.02	25.94	51.41	27.4	59.0
9	252.96	114.64	142.68	42.57	25.97	50.99	27.4	59.0
10	247.98	113.85	142.03	42.68	25.84	51.53	27.3	59.1

*c.c is the correlation coefficient

Table 2.13. Solutions from monomer searching from 25 to 5Å with 2.1Å *E. coli* and 2.2Å *P. putida*

2.4 Discussion

Pure KDPG aldolase, as judged by a single band on silver stained SDS-PAGE and IEF gels, was readily obtainable at a yield of 8mg L^{-1} in LB. Dynamic light scattering analysis of the protein in solution indicated it to be trimeric, although this was not conclusive. The protein readily crystallised in a variety of conditions; the optimum crystallisation conditions were found to be 20% PEG 6K, 0.075M citric acid pH 4.0, incubated at 293.5K, in sitting-drop vapour diffusion trays.

Crystals diffracted in-house to 2.8Å, although this data did not enable a molecular replacement solution to be found using the enzyme from *P. putida* firstly at 2.8Å resolution and then at 2.2Å using *AMoRe*. Crystals diffracted at the CLRC Daresbury synchrotron, PX9.6, to 2.17Å. It was thought that this much higher resolution data set would allow us to better search for a molecular replacement solution. However, we were still not able to find a solution and the question of whether KDPG aldolase from *E. coli* was dimeric or trimeric in solution remained.

A full discussion of why the molecular replacement techniques did not work will be given in Chapter 4.

Chapter Three

Structure determination of KDPG aldolase by multiwavelength anomalous dispersion methods

3.1 Abstract

KDPG aldolase crystals were previously obtained and data collected to 2.17Å. The cloning vector, pBluescriptII, was investigated as a potential source of selenium-enriched protein. Whilst it was possible to grow the construct in minimal media it would not grow in selenium-enriched minimal media.

The plasmid was re-cloned into the IPTG inducible pET-28b(+) vector which produces His-tagged protein. KDPG aldolase was over-expressed and purified in an identical manner to the pBluescriptII produced protein with one difference; the His-tag was removed by cleavage with thrombin overnight prior to the second chromatography step. This step was essential for successful crystallisation. The plasmid vector was transformed into the methionine auxotrophic cell line B834(DE3) and over-expressed in selenomethionine enriched minimal media. The Se-Met protein was purified in an identical manner to the native, incorporating the thrombin treatment. Crystals were obtained in the previously optimised conditions and found to be larger than their native counterparts.

Data were collected on fixed wavelength X-ray sources, both in-house and at CLRC Daresbury PX7.2. Data, when scaled against native 2.17Å data, were found to be non-isomorphous. MAD data sets were collected at ESRF Grenoble ID14-4 and CLRC Daresbury PX9.5, to 2.5Å and 3.0Å respectively. SOLVE was used to identify the positions of the selenium atoms. While an interpretable map was obtained for the ID14-4 data, no solution was obtained for the PX9.5 data.

3.2 Introduction

The crystallographic structure factor has two terms; a phase and an amplitude; if both are known the electron density which gave rise to the diffracted wave can be reconstituted by a Fourier synthesis. However, in crystallographic diffraction only the amplitude can be recorded whilst the phase is lost. The phase must then be obtained by other methods.

When the structure of a protein cannot be solved by molecular replacement the phases must be determined by *ab initio* methods. One increasingly common method is multiwavelength anomalous dispersion (MAD). Here the crystals are labelled by introducing anomalously scattering atoms. The location of the anomalous scatterer can be determined by analysis of the Patterson map. In cases where many atoms are present direct methods are usually required. The small variation in the real and imaginary components of the scattering factor of such an atom can be used to determine the phases of the native structure. Atoms which give rise to significant anomalous diffraction at experimentally useful incident X-ray wavelengths are advantageous in protein crystallography.

The majority of all MAD phased structures have used the methionine isostere selenomethionine to label the targeted protein (Deacon & Ealick, 1999). The incorporation of selenium into the host exploits the inability of the methionyl-tRNA synthetase of the host, for example *E. coli*, to distinguish between selenomethionine and methionine in the amino-acylation step of tRNA charging (Smith & Thompson, 1998). Several *E. coli* methionine auxotrophic strains have been developed including

DL41(DE3) and B834(DE3), which have lesions at their MetA genes. The B834(DE3) cell line utilises a T7 RNA polymerase promoter-based regulatory system and encodes an IPTG inducible chromosomal copy of the T7 RNA polymerase. Growth of the B834(DE3) cell line in complete media has also been shown to actually improve the expression levels of some native proteins.

It is not only the combination of selenium and methionine that can be used to obtain the desired anomalously scattering atoms. For example, a cysteine auxotrophic strain was used to incorporate selenocysteine into thioredoxin, with the subsequent formation of diselenide bridges. It is thought such diselenides may be more amenable to MAD studies due to enhanced resistance to oxidation.

There are several automated methods now available for solving crystal structures on the basis of the identification of the location of anomalous scattering atoms; *SOLVE* is one such program (Terwilliger & Berendzen, 1999). *SOLVE* will conduct all the steps of structure determination for MAD starting with raw intensities and ending with phases and a density map which can be viewed directly in *O* (Jones *et al.*, 1991). The ease of manipulation of *SOLVE* means it can run automatically for most investigations.

3.3 Experimental

Expression in Minimal Media

As mentioned in Chapter 2, KDPG aldolase was cloned into the pBluescriptII vector which is non-inducible, and as such it was unclear if the cells would be able to grow on a selenomethionine enriched minimal media. To investigate this 1L of methionine-enriched minimal media was grown at 310K, 175rpm for 3 hours. The cells were harvested by centrifugation; 8000 g, 15 mins, 277K and the pellets frozen at 253K. Immediately prior to purification the pellets were defrosted on ice and resuspended in ice-cold buffer, 20mM Tris-base pH 8.0, 1mM EDTA and 1mM PMSF, at a concentration of 10ml L^{-1} cell culture. The solution was sonicated on ice, 6 cycles of 30 s with 30 s rest period. The cell debris was removed by centrifugation; 18000 g, 15 mins, 277K. Analysis of the cell lysate by Coomassie stained SDS-PAGE showed a protein at 23kDa to be over-expressed.

The experiment was repeated using selenomethionine-enriched minimal media; all other experimental details were identical to the methionine-enriched protocol. Analysis of the cell lysate by Coomassie stained SDS-PAGE did not show any protein to be over-expressed. It was thus concluded the pBluescriptII vector was not amenable to growth on selenomethionine enriched minimal media.

pET Vector Cloning

To circumvent this problem *E. coli* KDPG aldolase cDNA was cloned into the pET-28b(+) vector by the Toone group, Duke University. The clone was transformed into BL21(DE3)pLysS in St Andrews. This vector contains a *lacZ* promoter and is thus

inducible on addition of IPTG. After induction the vector produces His-tagged protein; with the 6 histidines located at the N-terminus.

Initially a single colony was isolated from an Luria agar plus kanamycin plate and used to inoculate an overnight culture of 10ml LB plus 100μg ml^{-1} kanamycin and 50μg ml^{-1} chloramphenicol, incubated at 310K, 175 rpm shake. This culture was then used to inoculate 2L of Terrific Broth (TB) plus 100μg ml^{-1} kanamycin and 50μg ml^{-1} chloramphenicol and incubated at 310K, 175 rpm shake. The growth was monitored until an OD_{600} of 0.6 was reached, IPTG was then added to a final concentration of 0.4 mM and incubated for a further 3 hours. The cells were harvested by centrifugation 8000 g, 15 mins, 277K and the resultant pellet frozen at 253K. The pellet was defrosted on ice and resuspended in 20mM Tris-base, pH 8.0, 1mM EDTA, 1mM DTT at a concentration of 10ml L^{-1}. Following incubation at room temperature for 30 mins the viscosity of the solution was reduced by the addition of 20μg ml^{-1} DNAse I. To ensure complete cell lysis occurred the cells were sonicated on ice, 6 cycles of 30 s with 30 s rest. The cell debris was removed by centrifugation, 18000 g, 15 mins, 277K. The resultant supernatant was brought to 20% $(NH_4)_2SO_4$ at 277K and centrifuged to pellet insoluble proteins; 18000 g, 15 mins, 277K. This supernatant was dialysed against 20mM Tris-base, pH 8.0, to remove all salt prior to the initial purification step and filtered through a 0.2μm membrane.

This protein sample was applied to a POROS-HQ HPLC column, BioCad 700E, and eluted against an increasing salt gradient of 0 to 1000mM NaCl, 20mM Tris-base, pH 8.0. Coomassie stained SDS-PAGE analysis of the fractions showed a protein at 23kDa which was eluted at 100 to 120mM NaCl; this was identified as KDPG aldolase.

At this stage the protein sample was split into two equal parts to analyse the effect of the His-tag on the proteins' ability to crystallisation. One half was immediately subjected to the second purification step. It was brought to 30% $(NH_4)_2SO_4$ saturation at 277K. This was applied to a POROS high density phenyl HPLC column, BioCad 700E, previously equilibrated in 30% $(NH_4)_2SO_4$, 20mM Na_2HPO_4, pH 8.0. The protein was eluted in a decreasing salt gradient, 30 to 0% $(NH_4)_2SO_4$. Silver stained SDS-PAGE analysis of the fractions showed KDPG aldolase to be eluted at 10% $(NH_4)_2SO_4$ and to be present as a single band. The fractions corresponding to this peak were combined and concentrated under nitrogen pressure in an Amicon ultrafiltration unit at 277K.

The second half of the KDPG aldolase sample was digested with thrombin, to cleave the His-tag from the protein, at a concentration of 10units mg^{-1} fusion protein for 16 hours at 295K. This sample was then treated identically to the non-cleaved half. To ensure cleavage had only occurred at the His-tag and the proteins N-terminus was fully intact N-terminal sequencing was performed to confirm the integrity of the sample.

Crystallisation trials using KDPG aldolase samples with, and without, the tag were undertaken with Hampton screens I and II, incubated at 293.5K and 277K. Crystallisation trials using previously optimised conditions were also performed at 293.5K in both hanging and sitting drop plates. Following examination of the plates over a period of 2 weeks it was found that only the cleaved protein gave crystals; the His-tagged protein did not give any crystalline material under any of the conditions tested. At this stage it was decided to cleave the His-tag with thrombin after the first stage of purification in all future protein preparations.

Selenomethionine Enriched Protein Preparation

To ensure the protein from minimal media preparation would crystallise it was purified as before, with thrombin cleavage prior to the hydrophobic column. The recipe of this minimal media may be found in table 3.01 (Budisa *et al.*, 1995; Budisa *et al.*, 1997).

The pET-28b(+) vector containing the KDPG aldolase cDNA was transformed into the methionine auxotrophic *E. coli* strain B834(DE3). Initially this strain was grown on LB agar plates plus kanamycin. From this overnight cultures were grown in methionine enriched minimal media at 310K, 175 rpm and these cultures used to inoculate 2L of methionine enriched minimal media. The culture was grown until an OD_{600} of 0.6 was reached; this took approximately 24 hours. Over-expression was then induced by addition of IPTG to a final concentration of 0.4mM. 20μg ml^{-1} lysozyme was added to the cells which were then harvested by centrifugation, 8000 g, 15 mins, 277K and the resultant pellet frozen at 253K.

The pellet was defrosted on ice and resuspended in 20mM Tris-base, pH 8.0, 1mM EDTA, 1mM DTT at a concentration of 10ml L^{-1}. Coomassie stained SDS-PAGE analysis of the resultant compared with samples taken prior to IPTG induction showed a band at 23kDa, corresponding to KDPG aldolase, to be over-expressed.

Ingredient	Volume Required (ml L^{-1})
H_2O	800
1 M $(NH_4)_2SO_4$	7.5
1 M NaCl	8.5
1 M KH_2PO_4	55
1 M K_2HPO_4	100
1 M Glucose	20
1 M $MgSO_4$	1
1g L^{-1} CaCl	1
1g L^{-1} $FeSO_4$	1
μ-Elements ($ZnSO_4$, $MnCl_2$, $CuCl_2$)	2
5g L^{-1} Biotin	4
1g L^{-1} Thiamine	1
Methionine	58 mg

Table 3.01. Composition of methionine enriched minimal media

Crystal trials were performed under the previously optimised conditions. At this time Hampton additive screens were also performed using the optimised mother liquor conditions of 20% PEG 6K, 0.075M citric acid, pH 4.0. Further optimisation of additive screens selected 30% sucrose as increasing the quality of the crystals. Further trials were performed using 2μl protein plus 0.5μl additive and 2.5μl mother liquor.

It was decided to proceed with the growth of KDPG aldolase in selenomethionine enriched minimal media. The composition of this media was identical to that found in table 3.01 with selenomethionine replacing methionine. Growth was initiated with 10ml cultures of this media, incubated at 310K, 175 rpm for 24 hours. These cultures were then used to inoculate 2L of the same media and this was incubated at 310K, 175 rpm, until the OD_{600} 0.6 was reached; this took approximately 36 hours. Over-expression was then induced by addition of IPTG to a final concentration of 0.4mM and incubated for a further 12 hours. The cells were harvested by centrifugation, 8000 g, 15 mins, 277K and the resultant pellet frozen at 253K. The pellet was defrosted on ice and resuspended in 20mM Tris-base, pH 8.0, 1mM EDTA, 4mM DTT at a concentration of 10ml L^{-1}.

The cells were lysed by sonification as previously described and the cell debris pelleted by centrifugation; 18000 g, 15 mins, 277K. The supernatant was 0.2μm filtered and purified as before, again incorporating thrombin cleavage prior to the second chromatography step. In this case though all buffers used during the purification had 4mM DTT added to ensure DTT was present at all times during the purification. This was to avoid the oxidation and degradation of the selenium's, a characteristic that has proved problematic whilst locating the selenium edge during data collection. The purity of the seleno-methionine protein was analysed by silver stained SDS-PAGE and was found to be a single band running at approximately 23kDa as shown in figure 3.01.

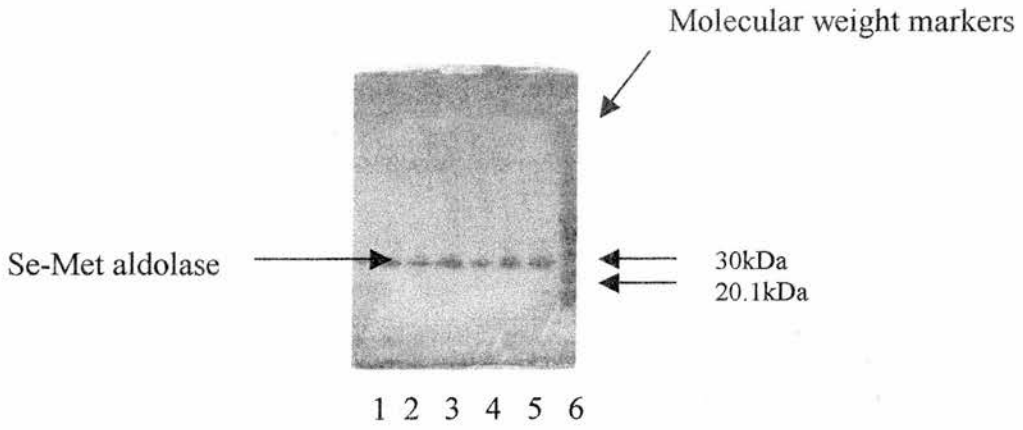


Figure 3.01. SDS PAGE gel of pure SeMet KDPG aldolase on a silver stained SDS-PAGE gel, lanes 1 to 5. Lane 6 is molecular weight markers.

To confirm the presence and incorporation of the selenium atoms into the protein MALDI-TOF spectroscopy was used, comparing pure native protein with pure selenomethionine protein. The weight difference between selenium and sulfur is 45Da, because there are 4 methionines in the KDPG aldolase sequence a weight difference of 180Da between native and seleno-methionine protein was expected. The traces for the two spectroscopy experiments are shown in figures 3.03 and 3.04. A difference of 186Da is clearly shown, indicating full incorporation of 4 seleniums. No intermediate peaks were visible, indicating none of the protein had only partial selenium incorporation, figure 3.05.

Crystals

The seleno-methionine protein was used in crystallisation experiments using the previously optimised conditions of 20% PEG 6K, 0.075M citric acid, pH 4.0, 30% sucrose with a protein concentration of 3mg ml^{-1}. Experiments were set up in both sitting and hanging drop trays. Protein crystals were obtained which were of much better quality than their native counterparts; they were larger in all three dimensions,

more stable with regard to cryo-protectants and generally more robust. A typical Se-Met crystal may be seen in figure 3.02.

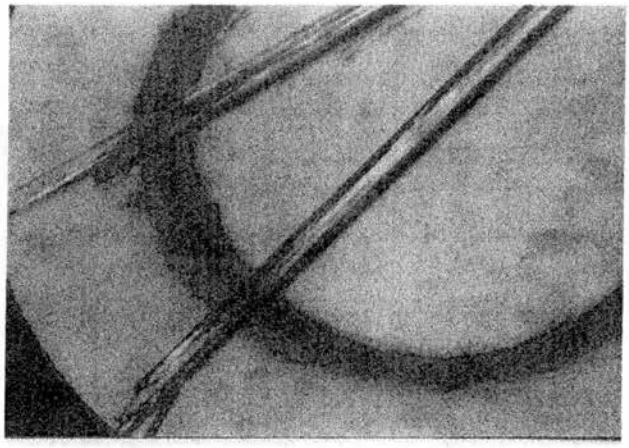


Figure 3.02. Se-Met KDPG aldolase sitting drop crystals

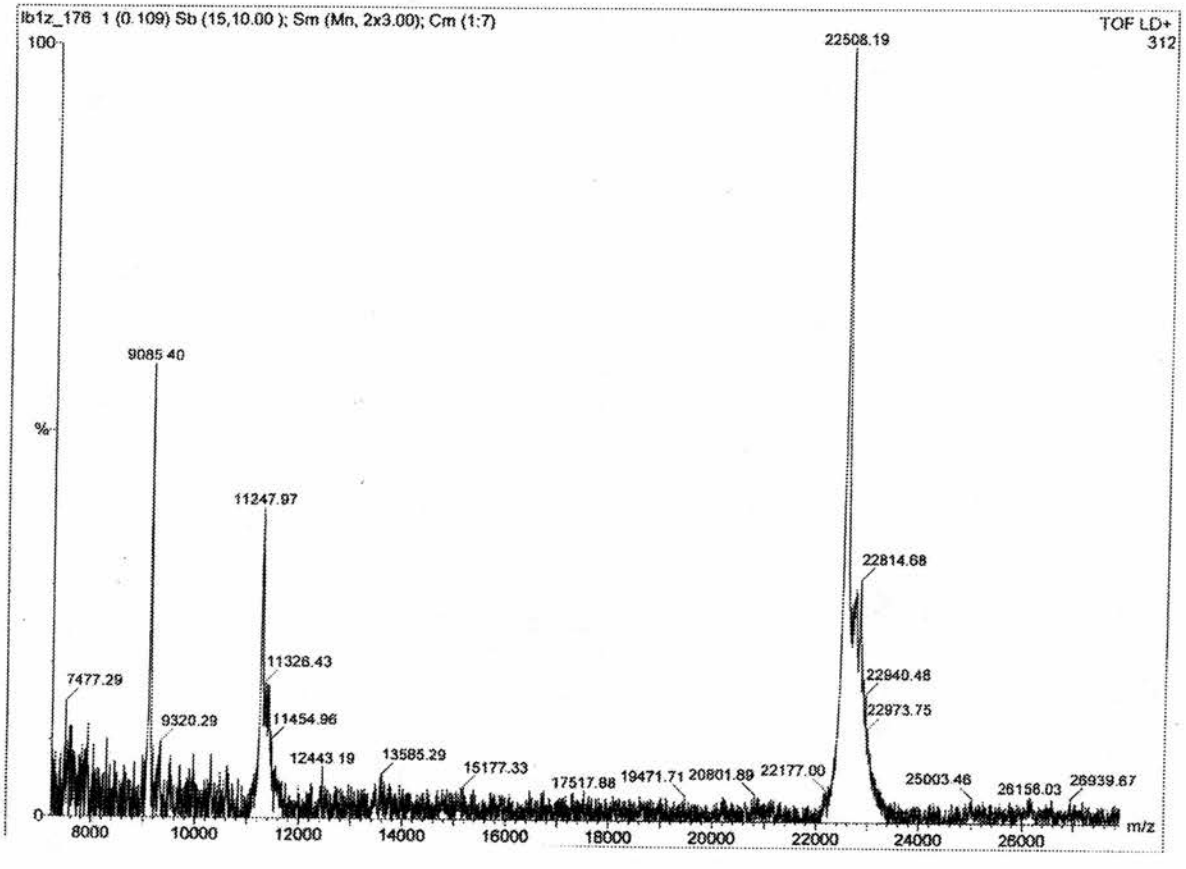


Figure 3.03. Native KDPG aldolase MALDI-TOF spectra.

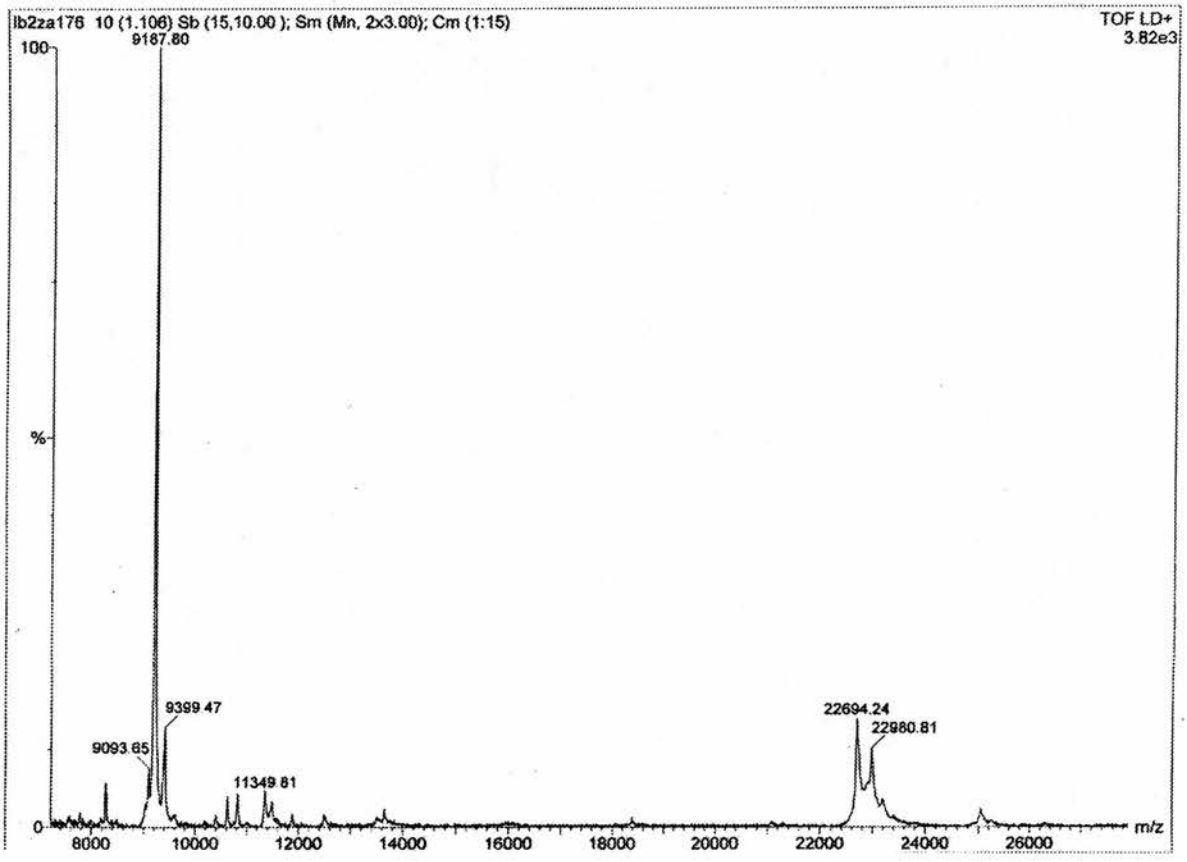


Figure 3.04. Se-Met KDPG aldolase MALDI-TOF spectra.

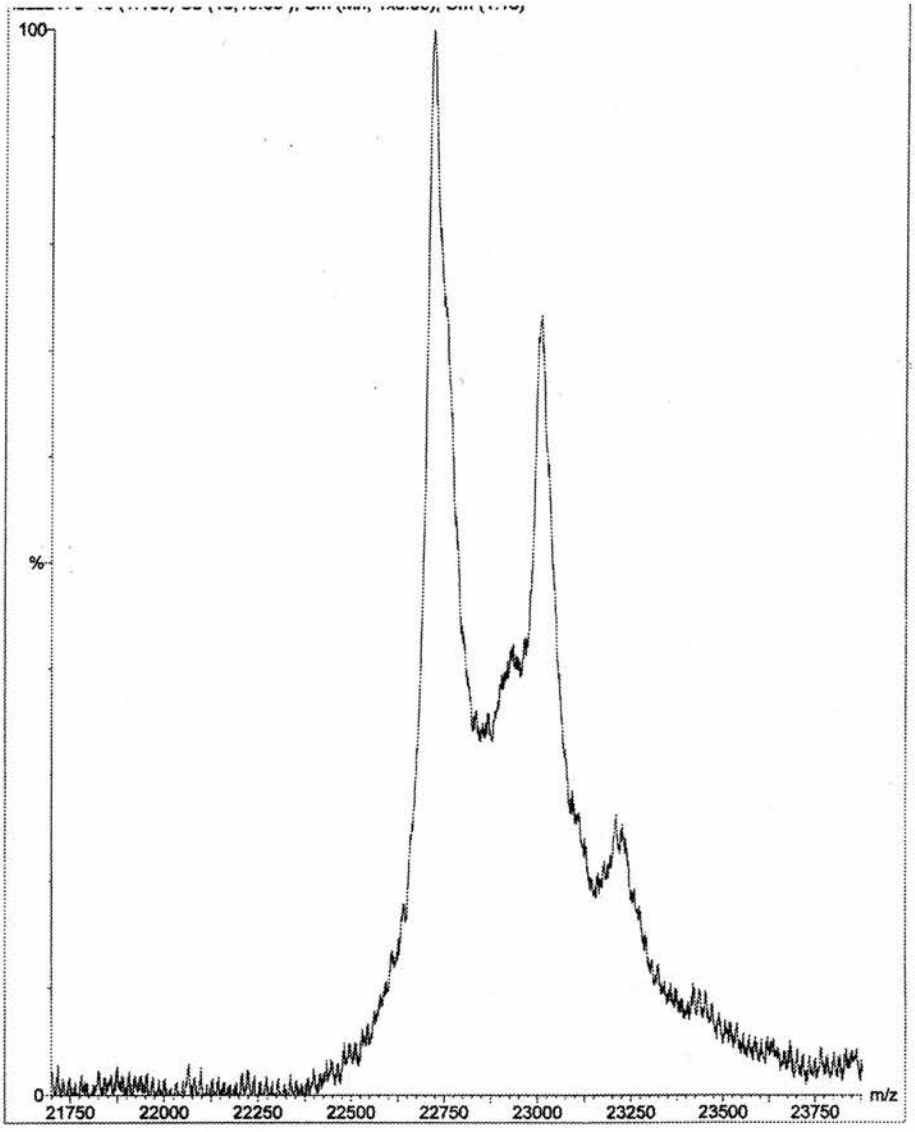


Figure 3.05. Expansion of the peak on the Se-Met KDPG aldolase MALDI-TOF spectra.

Data Collection

Copper Rotating Anode

A Se-Met crystal was soaked for 30 s in 15% glycerol prior to data collection at 130K. Data were collected from a single crystal; 97 images with an exposure time of 30 mins as 1° non-overlapping oscillations using a Nonius DIP2000 dual image plate. The data were indexed with *DENZO* and scaled with *SCALEPACK* (Otwinowski, 1996) to 2.9Å, with unit cell parameters a = 55.47Å, b = 83.97Å & c = 134.61Å, $\alpha = \beta = \gamma = 90°$. A complete summary of the data may be found in table 3.02.

Resolution (Å)	No. Reflections	Completeness (%)	R_{merge} (%)	Redundancy
30 – 6.23	1578	98.6	3.9	3.43
6.23 – 4.95	1503	99.5	6.2	3.70
4.95 – 4.33	1469	99.6	5.8	3.56
4.33 – 3.93	1442	97.1	6.2	3.18
3.93 – 3.65	1345	92.6	8.2	2.99
3.65 – 3.44	1264	86.6	11.0	3.00
3.44 – 3.27	1214	84.4	12.5	3.01
3.27 – 3.12	1182	81.8	15.1	2.97
3.12 – 3.0	1184	81.6	18.8	2.90
3.0 – 2.9	1164	81.7	21.8	2.79
30 – 2.9	13345	90.5	7.7	3.16

Table 3.02. 2.9Å Se-Met KDPG aldolase data statistics

These data were scaled against the 2.17Å native data using *SCALEIT* (CCP4, 1994). It was found that the derivative data were not completely isomorphous. The table 3.03

shows the statistical outputs from *SCALEIT*. For isomorphous data *diso* should decrease with resolution. Although a slight decrease is seen a more obvious decrease is expected.

Resolution (Å)	diso	R_{factor} (%)	Weighted-R (%)
11.0	35.0	11.5	12.6
8.5	40.5	11.3	13.0
7.2	35.5	13.1	13.6
6.3	29.2	13.5	12.5
5.7	31.8	15.8	14.2
5.3	36.8	17.5	16.2
4.9	36.7	16.0	14.7
4.6	39.8	15.8	15.7
4.4	39.8	15.6	15.4
4.1	41.4	16.8	16.4
4.0	41.3	16.2	15.4
3.8	42.9	18.2	17.4
3.7	36.8	17.0	16.3
3.5	36.1	18.5	17.4
3.4	36.8	19.9	18.0
3.3	32.2	18.3	15.8
3.2	32.7	20.3	17.6
3.1	32.4	21.3	18.1
3.0	28.9	21.5	18.4
11 – 3.0	35.9	17.2	15.8

Table 3.03. *SCALEIT* output statistics for Se-Met data scaled against native data

Daresbury PX7.2

Data were collected from a single Se-Met crystal on CLRC Daresbury station PX7.2 using the ADSC Quantum-4 CCD detector. The crystal was soaked for 30 s in 15%

glycerol prior to data collection at 100K. Data were recorded as 90 images with an exposure time of 60 s and an oscillation of 1° non-overlapping images, the crystal-to-detector distance was 210mm and the wavelength fixed at 1.488Å. Data were indexed and integrated with *DENZO* version 1.12d and merged with *SCALEPACK* (Otwinowski, 1996); the unit cell parameters were a = 55.80Å, b = 84.16Å & c = 135.32Å, $\alpha = \beta = \gamma$ =90°. Analysis of the data found the crystal was not isomorphous with the native data and thus the Patterson map could not be solved. It was decided to solve the structure by MAD.

Heavy Atom Derivatives

Whilst waiting for synchrotron time it was decided to search for heavy atom derivatives of the aldolase, using native crystals, in-house with the aim of obtaining isomorphous data. Solutions of heavy atoms were made up in the optimised crystallisation conditions, 20% PEG 6K, 0.075M citric acid pH 4.0. The heavy atom salts used were mercury acetate ($Hg_2(CH_3CO_2)_2$) mercury bromide (Hg_2Br_2) and platinum di-potassium hexa-chloride (K_2PtCl_6). Due to the insolubility of these heavy atom salts they were used as stock solutions for crystal soaking experiments. A drop of mother liquor plus heavy atom salt to a final concentration of between 1 and 10mM had a crystal transferred into it. The drop was prepared on either a cover slip or in a sitting drop plate depending on the original growth conditions.

The crystals were initially analysed under the microscope for evidence of changes in their appearance, for example cracking and re-annealing, which would indicate the protein had bound the heavy atoms. No major changes were observed for any of the salts at any of the concentrations. In addition to altering the concentration of the salts,

the time for which a crystal was soaked was also altered to an incubation time of between 2 and 48 hours. A number of these crystals were diffracted in-house at 130K after cryo-protection in mother liquor plus 15 % glycerol.

Data were collected as 30 images with an exposure time of 30 mins, as 1° non-overlapping oscillations using a Nonius DIP2000 dual image plate. The data were indexed with *DENZO*, scaled with *SCALEPACK* (Otwinowski, 1996) and converted into mtz format. The potential derivative data sets were then scaled against the native data. There were no obvious differences between any of the derivative data sets and the native, indicating the heavy atoms had not bound in the protein crystal, as is shown in table 3.04. From this it became obvious the structure would have to be solved by MAD using the selenomethionine derivative on a tunable beamline.

Heavy Atom Salt	Soak Duration	Data (Å)	Heavy Atoms Bound ?
$Hg_2(CH_3CO_2)_2$	1 hour	3.0	No
$Hg_2(CH_3CO_2)_2$	48 hours	2.9	No
Hg_2Br_2	4 hours	3.2	No
Hg_2Br_2	24 hours	3.0	No
K_2PtCl_6	24 hours	3.2	No

Table 3.04. A selection of the heavy atom derivatives analysed in-house

ID14-4 Data

Data were collected at ESRF Grenoble on the tuneable beamline ID14-4. An EXAFS scan was used to confirm the presence of the seleniums in the crystal and locate the edge. Three data sets were collected. One set of data was collected on the selenium edge at λ = 0.9794Å, one at the point of inflection on the selenium edge, λ = 0.9798Å, and a

third on the high-energy side of the selenium edge, $\lambda = 0.9393$Å. Data were collected as 108 images with an exposure time of 5 s, 1° non-overlapping images, with a crystal to detector distance of 250mm. Data were integrated with *DENZO* and scaled with *SCALEPACK* (Otwinowski, 1996) to 2.5Å, treating Bijovet pairs independently and gave cell constants of a = 55.06Å, b = 84.32Å, c = 134.32Å, $\alpha = \beta = \gamma = 90°$. Summary data for the three wavelengths is shown in table 3.05. Complete data analysis is shown in table 3.07. A typical Se-Met diffraction pattern is shown in figure 3.06.

	$\lambda = 0.9794$	$\lambda = 0.9798$	$\lambda = 0.9393$
Unique Reflections	22040	18160	27901
I/σI	21.9	21.4	23.0
Average Redundancy	2.0	2.1	1.98
Completeness (%)	92.7	93.7	88.2
Anom Complete (%)	84.4	84.3	87.4
R_{merge} (%)	5.4	6.3	5.5
Refined f'/f''	-9.24/5.54	-8.85/2.71	-2.25/3.35

Table 3.05. Summary of data from all three wavelengths on ID14-4

Selenium Positional Identification

The program *SOLVE* (Terwilliger & Berendzen, 1999) was used to identify the locations of the selenium's in the KDPG aldolase and calculate the phases. It identified 10 selenium sites, shown in table 3.06. The Z score for the sites found is 70.94, with a figure of merit value of 0.54. At this stage it was still unclear if the protein existed as a dimer or trimer. As there are 4 methionines in the sequence and 10 sites found it was deduced that the protein was trimeric, with 2 of the selenium's being disordered and thus not located by *SOLVE*.

The sites were not identified via the NCS symmetry matrix, calculated using *NCSFIND* (CCP4, 1994); it found only a 2-fold symmetry axis and not the true 3-fold. The non-crystallographic relationship was determined in *O* (Jones *et al.*, 1991) by manual inspection of the selenium positions. The seleniums were grouped into 3 groups of 3, with 1 group of 3 present in each monomer. The tenth site was found to be ordered in only one monomer. A point lying on the three-fold axis, approximately co-planar with the group of 3 seleniums was chosen as the molecular centre from which a mask of 40Å was calculated. This mask was used for the modification of the initial *SOLVE* map using the NCS averaging option of *DM* combined with solvent flattening and histogram matching, (Cowtan, 1994). The resulting map was readily interpretable; without averaging the map was non-interpretable. Examples of the maps from the initial *SOLVE* solution and after modification in DM are shown in figures 3.06 and 3.07. Full details of the structure tracing and refinement may be found in Chapter 4.

Site	x	y	z	Occupancy	B	Height / σ*
1	0.58	0.168	0.204	0.641	24.15	18.8
2	0.150	0.362	0.170	0.658	19.51	15.4
3	0.222	0.118	0.088	0.955	39.04	16.6
4	0.280	0.161	0.054	0.821	29.94	15.4
5	0.479	0.521	0.102	0.869	58.73	10.5
6	0.047	0.092	0.199	0.801	42.99	16.2
7	0.201	0.324	0.130	0.714	27.07	14.4
8	0.973	0.665	0.128	0.451	39.00	9.1
9	0.717	0.235	0.004	0.359	23.83	8.6
10	0.410	0.882	0.061	0.571	36.60	8.7

*The peak height is height of peak at this position divided by the r.m.s. of the map.

Table 3.06. Results from *SOLVE* ran using the ID14-4 data

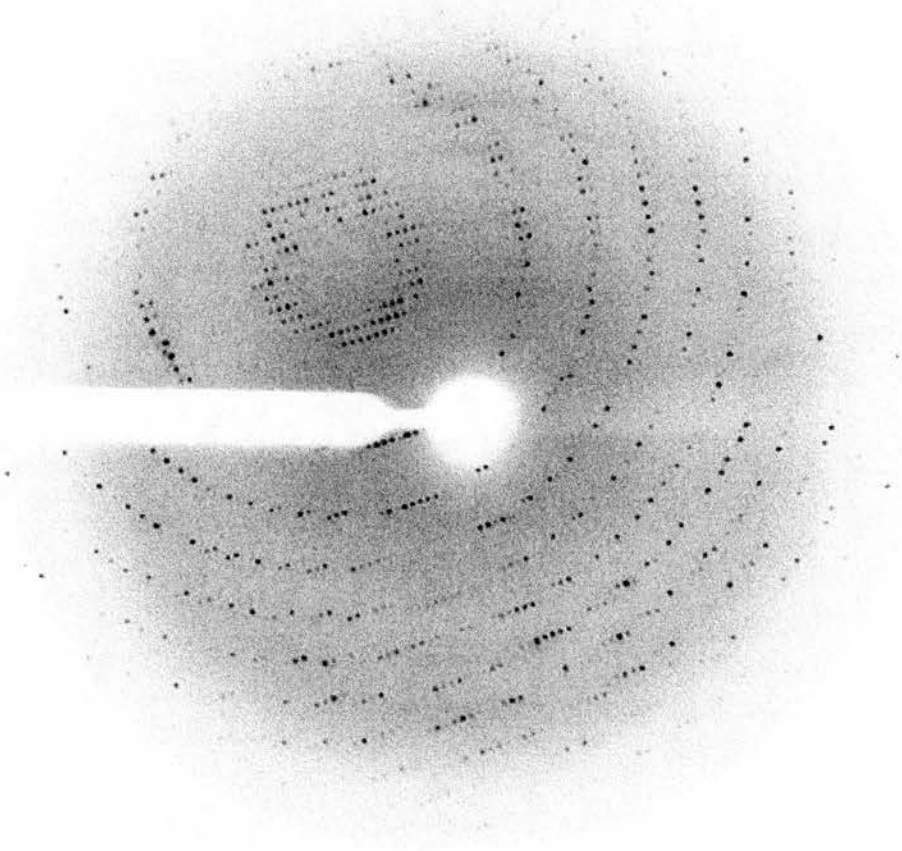


Figure 3.06. Se-Met KDPG aldolase diffraction pattern on ID14-4

$\lambda = 0.9794$Å

Resolution (Å)	No. Reflections	Completeness (%)	R_{merge} (%)	Redundancy
30.0 – 6.15	2662	95.0	6.5	2.15
6.15 – 4.89	2656	95.1	5.1	2.16
4.89 – 4.27	2676	96.2	4.8	2.17
4.27 – 3.88	2651	95.7	4.5	2.17
3.88 – 3.60	2717	96.8	4.5	2.89
3.60 – 3.39	2694	97.5	5.2	2.20
3.39 – 3.22	2684	97.8	5.4	2.18
3.22 – 3.08	2750	97.6	5.7	2.14
3.08 – 2.96	2678	95.7	5.8	2.11
2.96 – 2.86	2647	95.1	6.6	2.10
2.86 – 2.77	2674	96.4	7.8	2.07
2.77 – 2.69	2682	95.7	8.5	1.88
2.69 – 2.62	2400	87.7	9.0	1.66
2.62 – 2.56	2225	79. 2	9.4	1.52
2.56 – 2.50	1911	69.1	10.0	1.46
30 – 2.50	38707	92.7	5.4	2.02

$\lambda = 0.9798$Å

Resolution (Å)	No. Reflections	Completeness (%)	R_{merge} (%)	Redundancy
30.0 – 6.15	2661	94.6	5.7	2.26
6.15 – 4.89	2636	93.4	5.4	2.25
4.89 – 4.27	2644	94.6	5.1	2.25
4.27 – 3.88	2630	94.8	5.3	2.24
3.88 – 3.60	2690	95.3	5.5	2.24
3.60 – 3.39	2724	97.5	6.9	2.29
3.39 – 3.22	2729	98.2	7.0	2.29
3.22 – 3.08	2777	98.1	7.6	2.28
3.08 – 2.96	2699	96.6	7.7	2.25
2.96 – 2.86	2734	97.4	9.0	2.24
2.86 – 2.77	2738	98.1	10.4	2.23
2.77 – 2.69	2742	97.9	11.7	1.97
2.69 – 2.62	2554	91.1	13.1	1.73
2.62 – 2.56	2339	83.0	13.3	1.56
2.56 – 2.50	2086	75.3	15.1	1.50
30 – 2.50	39383	93.7	6.3	2.12

$\lambda = 0.9393$Å

Resolution (Å)	No. Reflections	Completeness (%)	R_{merge} (%)	Redundancy
30.0 – 5.65	3353	93.1	6.2	2.16
5.65 – 4.50	3376	93.8	5.2	2.20
4.50 – 3.93	3345	94.3	4.8	2.19
3.93 – 3.57	3430	94.9	4.9	2.18
3.57 – 3.32	3412	95.7	5.1	2.18
3.32 – 3.12	3503	97.6	5.4	2.18
3.12 – 2.96	3510	97.7	5.9	2.16
2.96 – 2.84	3468	96.8	6.7	2.13
2.84 – 2.73	3347	94.5	6.3	2.08
2.73 – 2.63	3382	94.5	7.6	2.04
2.63 – 2.55	3366	92.6	8.2	1.50
2.55 – 2.48	2949	83.6	9.6	1.60
2.48 – 2.41	2617	72.9	9.6	1.48
2.41 – 2.35	2328	64.4	11.0	1.42
2.35 – 2.30	2000	56.5	10.6	1.34
30 – 2.3	47386	88.2	5.5	1.98

Table 3.07. Statistical analysis of data from ID14-4

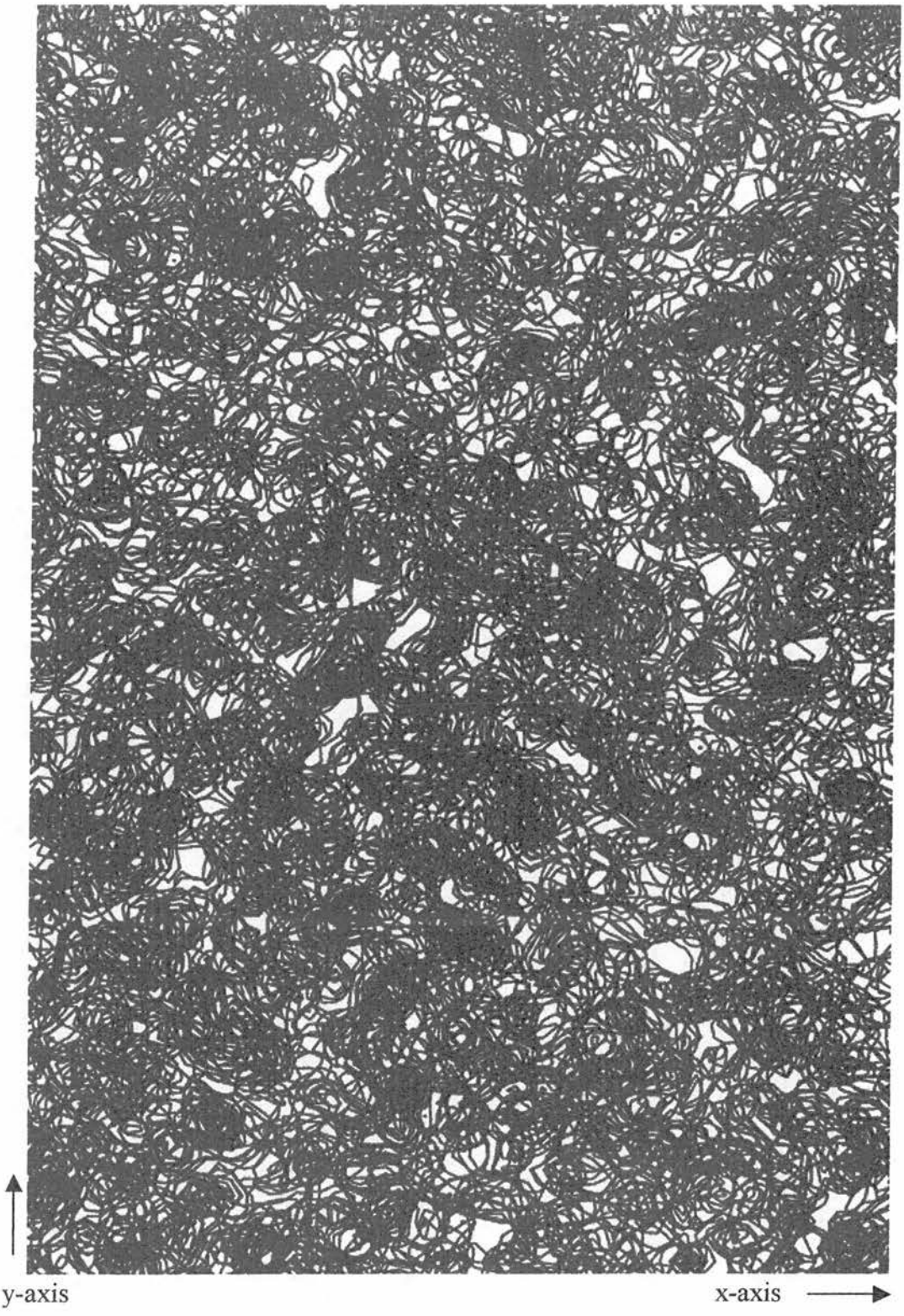


Figure 3.07. Density maps from *SOLVE* at 2.5Å. This map is clearly non-interpretable. For direct comparison this and figure 3.08 were taken over the same section of three-dimensional space.

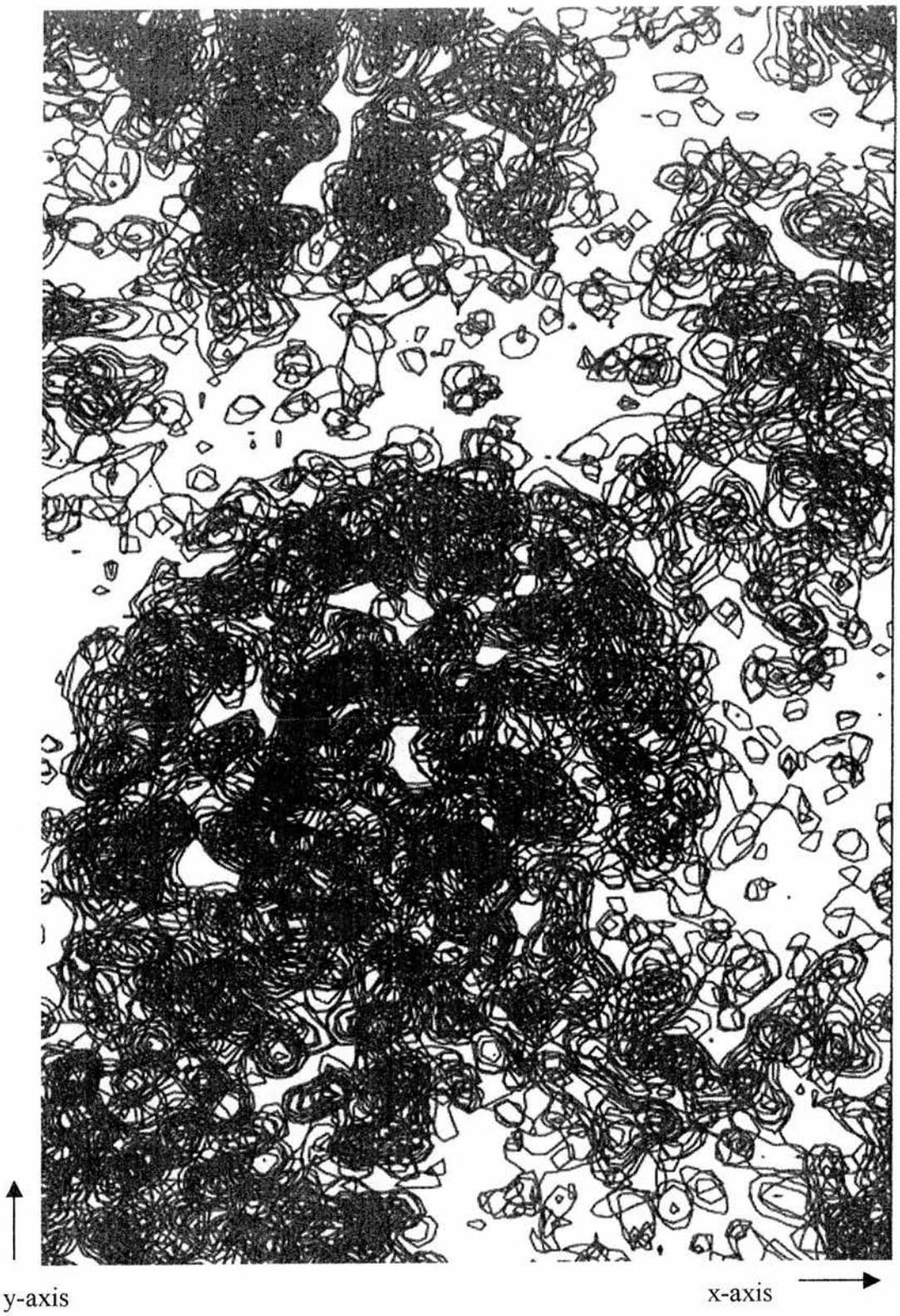


Figure 3.08. Density maps after solvent flattening and three-fold averaging in *DM* to 2.5Å. This map is clearly interpretable.

PX9.5 Data

Data were collected on the CLRC Daresbury synchrotron tuneable beamline PX9.5. An EXAFS scan was used to detect the presence of the selenium's in the crystal and locate the edge. Three data sets were collected; one set of data was collected on the selenium edge at λ = 0.797Å, one at the point of inflection on the selenium edge, λ = 0.9783Å, and a third on the high-energy side of the selenium edge, λ = 0.9226Å. Data were collected as 90 images with an exposure time of 180 s, 1° non-overlapping images. Data were integrated with *MOSFLM*, scaled with *SCALA* (Leslie, 1992) to 3.0Å and gave cell constants of a = 54.92Å, b = 84.39Å, c = 134.20Å, $\alpha = \beta = \gamma = 90°$. Summary data for the three wavelengths is shown in table 3.08. Complete data analysis is shown in tables 3.10.

	$\lambda = 0.9797$	$\lambda = 0.9783$	$\lambda = 0.9226$
Unique Reflections	11456	11946	8583
I/σI	3.8	6.8	6.2
Average Redundancy	3.6	4.6	4.3
Completeness (%)	97.4	98.5	74.4
Anom Complete (%)	88.2	95.2	66.6
R_{merge} (%)	13.7	9.2	10.9
Refined f'/f'	1.99/2.04	-11.48/5.33	-15.79/4.33

Table 3.08. Summary of data from all three wavelengths on PX9.5

Selenium Positional Identification

The program *SOLVE* (Terwilliger & Berendzen, 1999) was used to solve the structure of the KDPG aldolase. It identified 8 out of the theoretical 12 selenium sites; the solutions are shown in table 3.09. The Z score for the sites found was 9.00 with a figure of merit value of 0.38. From the data collected on ID14-4 it was known the protein was

trimeric rather than dimeric, thus the missing seleniums are a function of the quality of the data collected on PX9.5 and not a true reflection of the proteins structure.

Site	x	y	z	Occupancy	B	Height / σ*
1	-0.173	-0.497	-0.117	1.172	35.1	4.3
2	0.362	0.957	0.177	1.279	47.8	3.1
3	0.267	0.178	0.159	1.917	29.7	5.4
4	0.212	0.226	0.170	0.745	15.0	3.2
5	0.881	0.690	0.052	0.203	15.0	4.1
6	0.365	0.089	0.117	1.742	51.9	4.3
7	0.072	0.366	0.075	1.092	19.2	4.2
8	0.968	0.729	0.073	0.098	15.0	3.4

*The peak height is height of peak at this position divided by the r.m.s. of the map.

Table 3.09. Output from *SOLVE* ran using PX9.5 data

$\lambda = 0.9797$Å

Resolution (Å)	No. Reflections	Completeness (%)	R_{merge} (%)	Redundancy
70.0 – 9.52	1569	96.7	11.7	3.4
9.52 – 6.73	2827	97.8	9.2	3.7
6.73 – 5.50	3695	98.6	8.1	3.7
5.50 – 4.76	4369	98.8	7.6	3.7
4.76 – 4.26	4903	99.0	6.3	3.8
4.26 – 3.89	5416	99.1	5.8	3.8
3.89 – 3.60	5881	99.2	5.5	3.8
3.60 – 3.37	6219	99.3	4.6	3.8
3.37 – 3.17	6643	99.4	6.0	3.8
3.17 – 3.0	4144	97.4	3.7	2.7
70 – 3.0	45666	97.4	13.7	3.6

$\lambda = 0.9783$Å

Resolution (Å)	No. Reflections	Completeness (%)	R_{merge} (%)	Redundancy
70.0 – 9.54	1981	98.2	5.6	4.2
9.54 – 6.74	3585	98.7	5.8	4.6
6.74 – 5.51	4671	99.0	7.8	4.8
5.51 – 4.77	5602	99.2	7.5	4.8
4.77 – 4.27	6285	99.3	7.4	4.9
4.27 – 3.89	6901	99.5	8.4	4.9
3.89 – 3.61	7531	99.6	10.6	4.9
3.61 – 3.37	7909	99.6	14.0	4.9
3.37 – 3.18	8401	99.7	17.0	4.8
3.18 – 3.02	5684	99.5	23.9	3.4
70 – 3.02	58559	98.5	9.2	4.6

$\lambda = 0.9226$Å

Resolution (Å)	No. Reflections	Completeness (%)	R_{merge} (%)	Redundancy
70.0 – 9.49	1331	68.6	4.8	4.1
9.49 – 6.71	2427	69.9	4.9	4.4
6.71 – 5.48	3183	70.7	8.3	4.5
5.48 – 4.74	3739	71.3	8.2	4.5
4.74 – 4.24	4209	72.0	8.1	4.4
4.24 – 3.87	4660	72.6	10.0	4.4
3.87 – 3.59	5132	73.1	13.5	4.4
3.59 – 3.35	5291	73.6	19.0	4.3
3.35 – 3.16	5573	74.0	22.4	4.1
3.16 – 3.0	5589	74.4	24.6	3.9
70 – 3.0	41134	74.4	10.9	4.3

Table 3.10. Statistical analysis of data from PX9.5

3.4 Results & Discussion

Se-Met Enriched Protein

To enable the production of selenomethionine enriched KDPG aldolase the plasmid was cloned into the IPTG inducible pET-28b(+) expression vector. This vector produces His-tagged protein. Over-expressed protein was purified in an analogous manner to that from the pBluescriptII vector; the only difference being a thrombin cleavage step to remove the His-tag prior to the second chromatography column. Silver stained SDS-PAGE analysis showed the protein to exist as a single band at approximately 23kDa. It was found that the protein with the tag would not crystallise.

The vector was transformed into the methionine auxotrophic *E. coli* strain B834(DE3) and grown on a minimal media. Protein purified from this source was amenable to crystallisation under previously optimised conditions. To produce the selenomethionine enriched KDPG aldolase the protein was over-expressed in a minimal media enriched with selenomethionine. Purification was identical to that for native protein expressed in the pET vector including thrombin cleavage. The yield of Se-Met KDPG aldolase was 2mg L^{-1}. The Se-Met protein was found to crystallise under the conditions optimised for the native protein; the crystals were found to be much larger and more robust when compared with the native ones.

Data Collection

Initially data were collected on fixed wavelength X-ray sources. Analysis of these data showed it to be non-isomorphous. Data from heavy atom soaks on native crystals at this time failed to identify an isomorphous derivative.

	ID14-4	PX9.5
Z score	70.94	9.0
Figure of Merit	0.54	0.38
No. Sites Found	10	8
Average B factor	34.1	28.5
Average Height / σ	13.7	4.0

Table 3.12. Comparison of *SOLVE* statistics for the two MAD data sets

The sites were identified for the ID14-4 data and the symmetry operators between the sites derived manually in *O*. For PX9.5 no convincing operators were found thus those derived from the ID14-4 data were used. A mask of 40Å was calculated at the midpoint of 9 of the ID14-4 sites. This mask was used to improve the initial *SOLVE* maps (these maps were non-interpretable in both cases) via the solvent flattening and NCS averaging options of *DM*. Tables 3.13 and 3.14 show a comparison of the non-crystallographic symmetry correlation's from *DM* for the two *SOLVE* solutions. It is clearly shown that the starting phase set from the PX9.5 data is too poor to continue working with.

	ID14-4 Data		PX9.5 Data	
	Initial c.c*	Final c.c	Initial c.c	Final c.c
Monomer 1	0.889	0.911	0.998	0.996
Monomer 2	0.262	0.845	0.068	0.225
Monomer 3	0.252	0.843	0.077	0.230

*c.c is the correlation coefficient

Table 3.13. Correlation coefficients for NCS operators on the trimer for each data set. The final values obtained for the ID14-4 data are typical for a solution; those obtained for PX9.5 are typical of no solution.

	ID14-4 Data		PX9.5 Data	
	Initial mean FOM's	Final mean FOM's	Initial mean FOM's	Final mean FOM's
Monomer 1	0.562	0.776	0.3925	0.462
Monomer 2	0.540	0.766	0.331	0.434
Monomer 3	0.544	0.717	0.279	0.332

Table 3.14. Figure of Merits for the trimer for the three highest resolution shells for each data set.

The initial and final figures of merit of the two phase sets are consistent with a solution for the ID14-1 data and no solution for the PX9.5 data. In conclusion the ID14-4 data led to a readily interpretable map that allowed the structure to be fully traced (Chapter 4). The PX9.5 data was not of sufficient quality to provide an interpretable map and thus could not be used to trace the structure of KDPG aldolase.

The reasons behind such a variation in data quality are not clear. The quality of the EXAFS scan on PX9.5 may not have been accurate enough to fully identify the peak and inflection wavelengths which were required to be collected. The beamline itself may have not been correctly tuned and thus not produced X-rays of the desired wavelength; or if it was correctly tuned it may have drifted during the course of data collection. The time required to collect all three wavelengths on PX9.5 was approximately 36 hours, the time to collect on ID14-4 was under 1 hour, thus some degree of the difference in the quality of the data may be attributed to the decay of both the beam and the crystal on PX9.5 compared with ID14-4. The user support and station set-up and overall organisation are much more tailored towards the inexperienced user at the ESRF.

The final conclusion is that the MAD data collected on ID14-4 were significantly better than that collected on PX9.5. It is important to note the Se-Met crystals were from the same preparation and were frozen and handled in an identical manner, thus the only notable variable was the quality of the beamline and facilities for EXAFS scans available at each synchrotron.

Chapter 4

Refinement & analysis of KDPG aldolase

4.1 Abstract

The initial model was traced into the solvent flattened and NCS averaged map. The model used to solve the 2.17Å native data by molecular replacement. This model was refined using CNS. A citrate anion from the crystallisation buffer was found bound in the active site. An R_{factor} of 19.5% and R_{free} of 25.1% were obtained for this 2.17Å data. The overall fold of the enzyme was found to be typical of the aldolase family; a TIM barrel fold. The enzyme is a trimer rather than the dimer or tetramer commonly associated with aldolase enzymes. Superposition of the P. putida enzyme against the E. coli enzyme found large differences between the two. This was attributed to errors in the P. putida structure rather than any genuine structural differences between the two enzymes.

4.2 Introduction

A unified mechanism of Class I aldolase enzymes has been proposed by two groups based on studies with transaldolase (Jia *et al.*, 1997) and FBP aldolase (Dalby *et al.*, 1999). This mechanism proposes a principle catalytic base and a secondary catalytic acid.

The proposed role of the acidic and basic residues implicated in the mechanism is not as strongly supported as that of the catalytic lysine. The mechanism for the transaldolase shares several features with the enzymatic mechanisms proposed for FBP aldolase (Dalby *et al.*, 1999). The two acidic residues in transaldolase, Asp17 and Glu96, are also found in FBP aldolase; Asp33 and Glu197. However, there is no additional lysine residue in the active site of transaldolase whilst there is in FBP aldolase.

This second lysine, Lys146 in FBP aldolase, is located close to the Schiff base forming lysine and has been implicated in the binding of the C1 of the phosphate group. It is possible that because the natural substrates of transaldolase are not phosphorylated at this position there is no requirement for this residue to be conserved. This second lysine was also proposed to be important in the formation of the Schiff base by acting as an acid / base catalyst. In transaldolase a conserved Thr33 is found where the second lysine is in FBP aldolase. While this residue's side chain can participate in hydrogen bonding and thus may stabilise the Schiff base intermediate, it is unlikely it acts as an acid / base catalyst for manipulation of the Schiff base intermediate.

Figure 4.01 shows an alignment of four aldolase enzymes known to operate under the Class I mechanism; FBP aldolase and KDPG aldolase from *E. coli*, KDPG aldolase from *Z. mobilis* and transaldolse from *E. coli*. The alignment demonstrates the relatively low similarity shown between them; this is more striking when it is considered they all have the same fold and, as mentioned, utilise a very similar mechanism. Figure 4.02 shows the mechanism of transaldolase (Jia *et al.*, 1997).

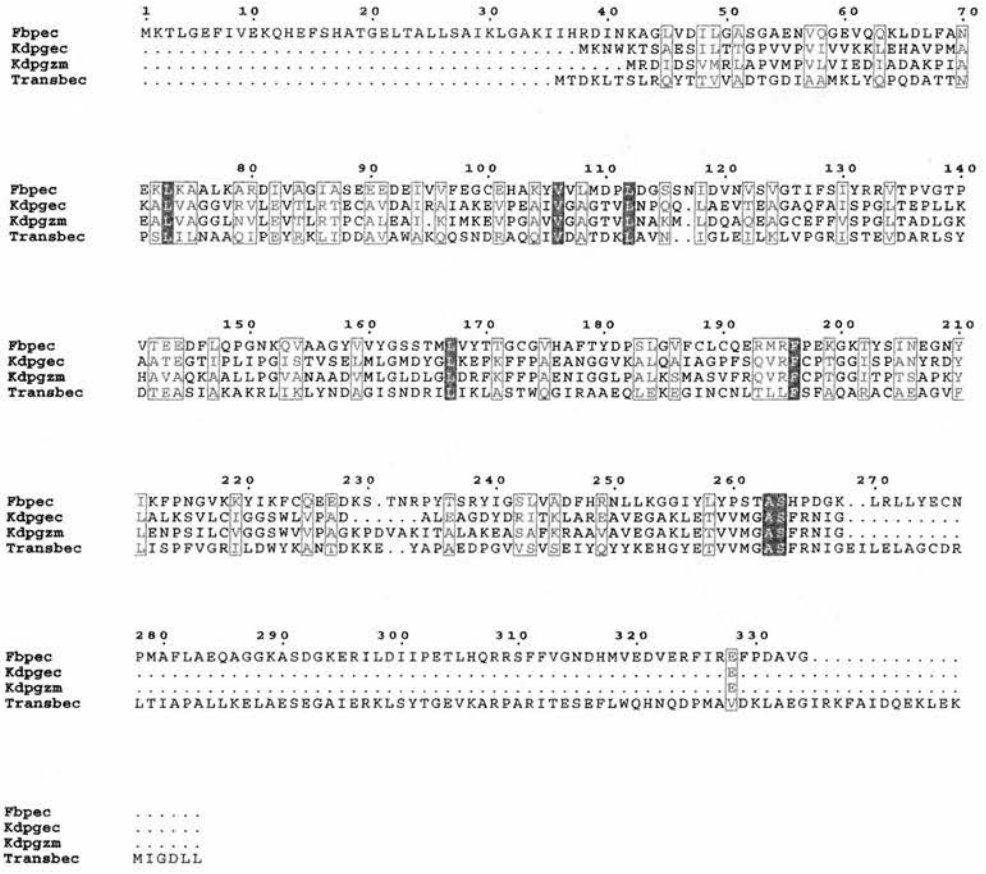


Figure 4.01 Sequence alignment of Class I aldolase enzymes using ESPript. From top to bottom; FBP aldolase from *E. coli*, KDPG aldolase from *E. coli*, KDPG aldolase from *Z. mobilis*, Transaldolase B from *E. coli*.

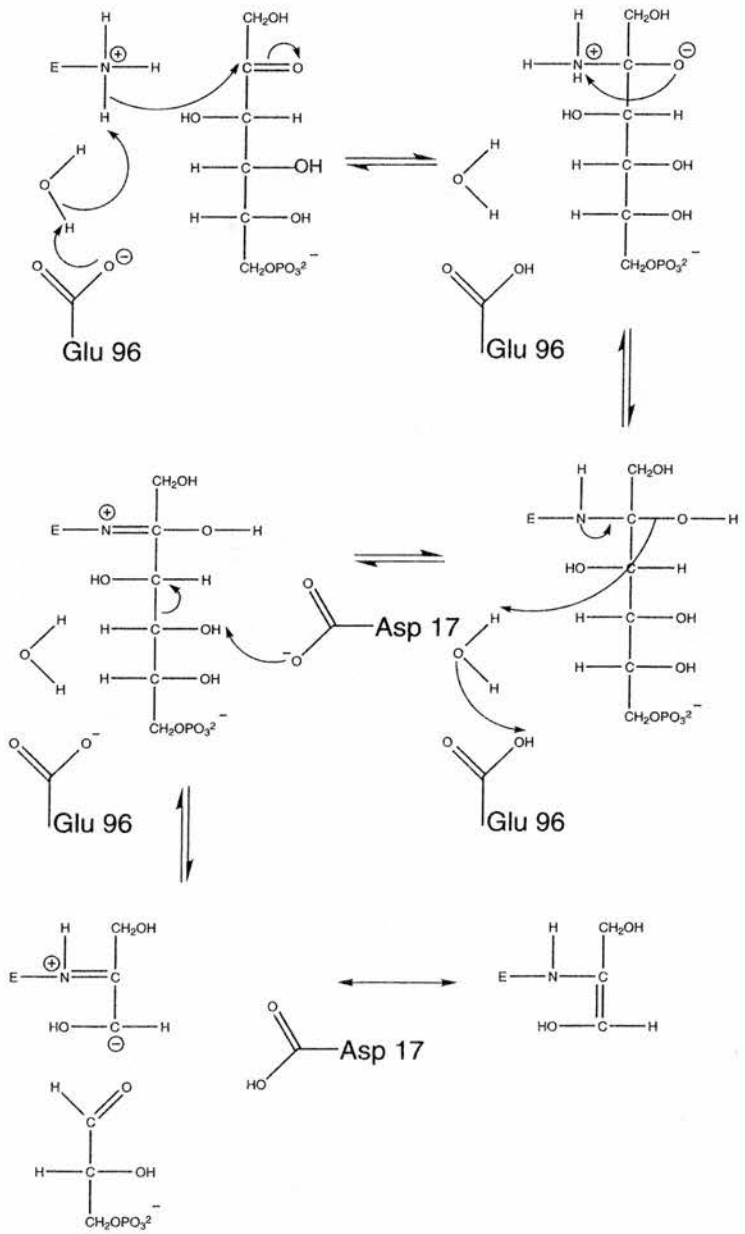


Figure 4.02. The proposed mechanism of transaldolase (Jia *et al.*, 1997).

4.3 Experimental

Initial Tracing of SeMet Maps

The *fo* map calculated to 2.5Å was viewed in *O* (Jones *et al.*, 1991). At this initial stage the tracing of the map using *O's* "bones" function was greatly facilitated by the identification of the selenium positions corresponding to the methionine residues 122 and 125. A tyrosine residue, at position 126, was clearly visible and thus allowed the determination of the direction in which to trace; forward from 125 and backwards from 122. In this manner it was possible to trace 63 residues forwards from 125 and 49 residues backward from 122 with certainty.

At this point it was decided to trace from the other 2 seleniums present. In only one monomer was the N-terminal methionine identified. This monomer was designated monomer A. From this residue the map was traced until a degree of ambiguity was located regarding the direction of the main chain. The final selenium was located, residue 32, and again the map traced. This strategy allowed a complete trace of all 213 residues to be made.

Once the initial tracing had been completed the map and alanine model were re-analysed and residues from the *E. coli* KDPG aldolase sequence assigned in place of the alanines. After manual manipulation in *O* to ensure all side chains were in density this monomer was used to generate the other two monomers via *LSQKAB* (Kabsch, 1976). After examining all 3 monomers, to ensure all side chains were in density, this model was used to find a solution to the 2.17Å native data using molecular replacement.

Molecular Replacement

The native structure of KDPG aldolase was determined using molecular replacement as implemented in the CCP4 program *AMoRe* with data from 15 to 3Å, with a search radius of 30Å (CCP4, 1994; Navaza & Saludjian, 1997). The trimeric *E. coli* KPDG aldolase model built from the Se-Met data was used as the search model. As may be seen from tables 4.01 to 4.03 the solutions are clearly related by 120°; the three-fold axis. All solutions from *AMoRe* are shown.

Solution	α	β	γ	c.c*	σ
1	169.84	55.42	187.75	21.6	12.7
2	169.89	56.53	64.54	20.8	12.2
3	168.65	55.88	305.58	18.4	10.8

Table 4.01. Solutions from rotation function

Sln	α	β	γ	x	y	z	c.c*	R_{factor}	σ
1	169.84	55.42	184.75	0.2195	0.3252	0.1460	48.6	46.9	14.4
2	169.99	56.53	64.54	0.2198	0.3253	0.1461	45.6	48.1	14.1
3	168.65	55.88	305.58	0.2190	0.3253	0.1464	40.9	50.0	12.4

Table 4.02. Solutions from translation function

Solution	α	β	γ	x	y	z	c.c*	R_{factor}
1	170.13	55.48	184.36	0.2195	0.3264	0.1458	59.4	39.1

*c.c is the correlation coefficient

Table 4.03. Solution from fitting function

The *2fo-fc* and *fo-fc* maps at 2.17Å were much clearer than the 2.5Å *fo* map. It was straightforward to identify the exact positions of all of the residues and their side chains. Examples of the maps are shown in figure 4.03. Whilst almost the entire alanine model had been correctly positioned some manual rebuilding of residue side chains was required; the details of which had not been clear with the lower resolution maps. Only

one section of the structure was not well fitted by the *2fo-fc* map; the external loop consisting of residues 188 to 196. At this point all residues were carefully re-examined and side chains not in density had their occupancy set to zero for refinement.

Refinement

Prior to commencing refinement 5% of the data were removed from all further calculations to give a free-R_{factor} from which the quality of refinement could be monitored. The free-R_{factor} does not validate the structure; it validates the refinement protocol. The refinement program Crystallography and NMR Systems (CNS) was used to refine the structure (Brunger *et al.*, 1998). For each round of positional refinement 300 cycles were run; for each round of temperature factor refinement 30 cycles were run. Initially the model was subjected to two rounds of positional and one round of temperature factor refinement. The model was next subjected to torsional dynamics, with a starting temperature of 5000K and a stepwise slow cool rate of 25K to 300K. Torsional dynamics was run 5 times using different initial velocities; from these five models the one with the lowest R_{factor} was selected.

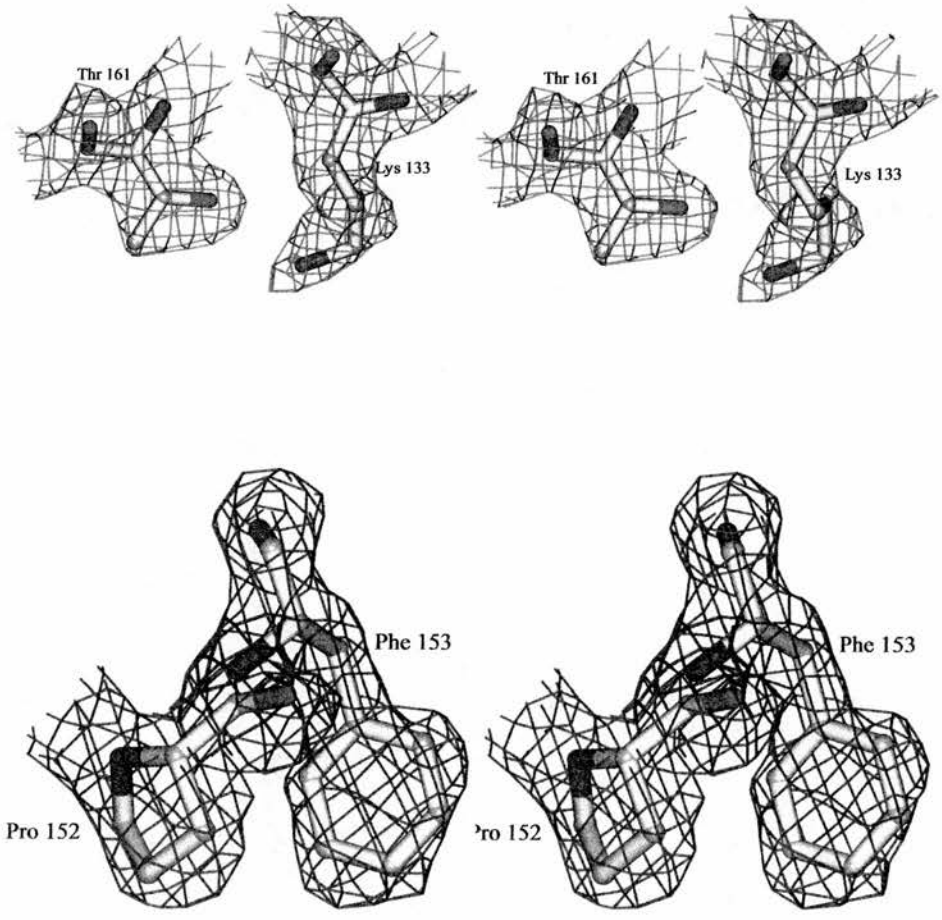


Figure 4.03. Electron density *2fo-fc* maps shown in stereo, contoured at 1σ for the 2.17Å data

The model was adjusted in *O*, subjected to further automated refinement and water atoms were added in *CNS*. Water-pick was used to select water atoms which satisfied the following conditions; they corresponded to a peak of greater than 3σ in the *fo-fc* map and made hydrogen bonds to protein atoms or other water molecules at a distance of not less than 2.4Å and not more than 3.5Å. When subjected to B-factor refinement waters with a B-factor of 70 or higher were removed. Once the waters were added they were subjected to further rounds of refinement and examined manually in *O*; any found not to be in the *2fo-fc* density map, contoured at 1σ, were removed.

Three water molecules were found to have been placed in connected density in the same position in all three monomers. Further analysis showed this density to be located within hydrogen bonding distance of the Schiff base forming lysine, residue 133. The waters in this position in all three monomers were removed and the structure subjected to further automated refinement. Analysis of the *fo-fc* map in *O* showed defined molecular density to be present in the active site. This was identified as a citrate ion which is used in the crystallisation of KDPG aldolase. A citric acid molecule was obtained from the Hetero-Compound Information centre Uppsala (HIC-Up) and modelled into the *fo-fc* density in each monomer. The citrate refined to an almost identical orientation in all three monomers

At this stage the loop region, comprising residues 188 to 196, in all three monomers remained ill defined. Omit maps were generated and used in *O* to more accurately trace this loop. To do this the residues in the loop had their occupancy set to zero and the model was subjected to simulated annealing. After a further round each of positional refinement and temperature factor refinement, an *fo-fc* map was generated. This map

was used to position the residues in the loop's main chain into density. Some ambiguity remained with respect to the side chains.

Prior to the last step the model was examined to ensure side chains which had come into density had their occupancy set to 1. A final round of automated refinement was performed, the statistical analysis of which may be found in table 4.04.

Statistic	KDPG aldolase to 2.17Å
R_{factor}●	19.5%
R_{free}	25.1%
R_{factor} (2.24 -2.17Å)	25.6%
R_{free} (2.24-2.17Å)	33.6%
Average B-factor■	19Å^2
RMSD* bonds	0.017Å
RMSD* angles	1.86°
B-factor Deviation bonds■	1.15Å^2
B-factor Deviation angles■	1.71Å^2
Residues in most favoured region of Ramachandran plot	92.9%
Number of non-hydrogen atoms	5246
Number of atoms with occupancy set to zero	186

*RMSD calculated via *LSQMAN* ●The difference between the absolute value of the observed structure factor amplitudes from the data & those calculated from the model, for all reflections, divided by the sum of the observed structure factor amplitudes ■B-factor deviation for bonded atoms

Table 4.04. Statistical analysis of the final KDPG aldolase model

Procheck was run as part of CCP4i (CCP4, 1994). The statistics from the Ramachandran plot (Ramakrishnan & Ramachandran, 1965) are given in table 4.05. The plot is shown in figure 4.04.

Statistic	No. of Residues	% of Residues
Residues in most favoured regions	484	92.9
Residues in additional allowed residues	37	7.1
Residues in generously allowed regions	0	0
Residues in disallowed regions	0	0

Table 4.05. Statistical analysis of the Ramachandran plot for KDPG aldolase

In the Ramachandran plot glycine residues are represented by a filled triangle; non-glycine residues are represented by a filled square. Neither Proline or Glycine residues are restricted to any defined region of the plot. The more red the region the more favourable the φ -ψ combination. As may be seen from table 4.05 and the plot, figure 4.03, no residues lie in either the generously allowed or the disallowed regions. In the plot the regions are labelled as shown in table 4.06.

Label	Definition
A	Core α
a	Allowed α
~a	Generous α
B	Core β
b	Allowed β
~b	Generous β
L	Core left handed α
l	Allowed left handed α
~l	Generous left handed α
p	Allowed ψ
~p	Allowed ψ

Table 4.06. Terms used in the Ramachandran plot

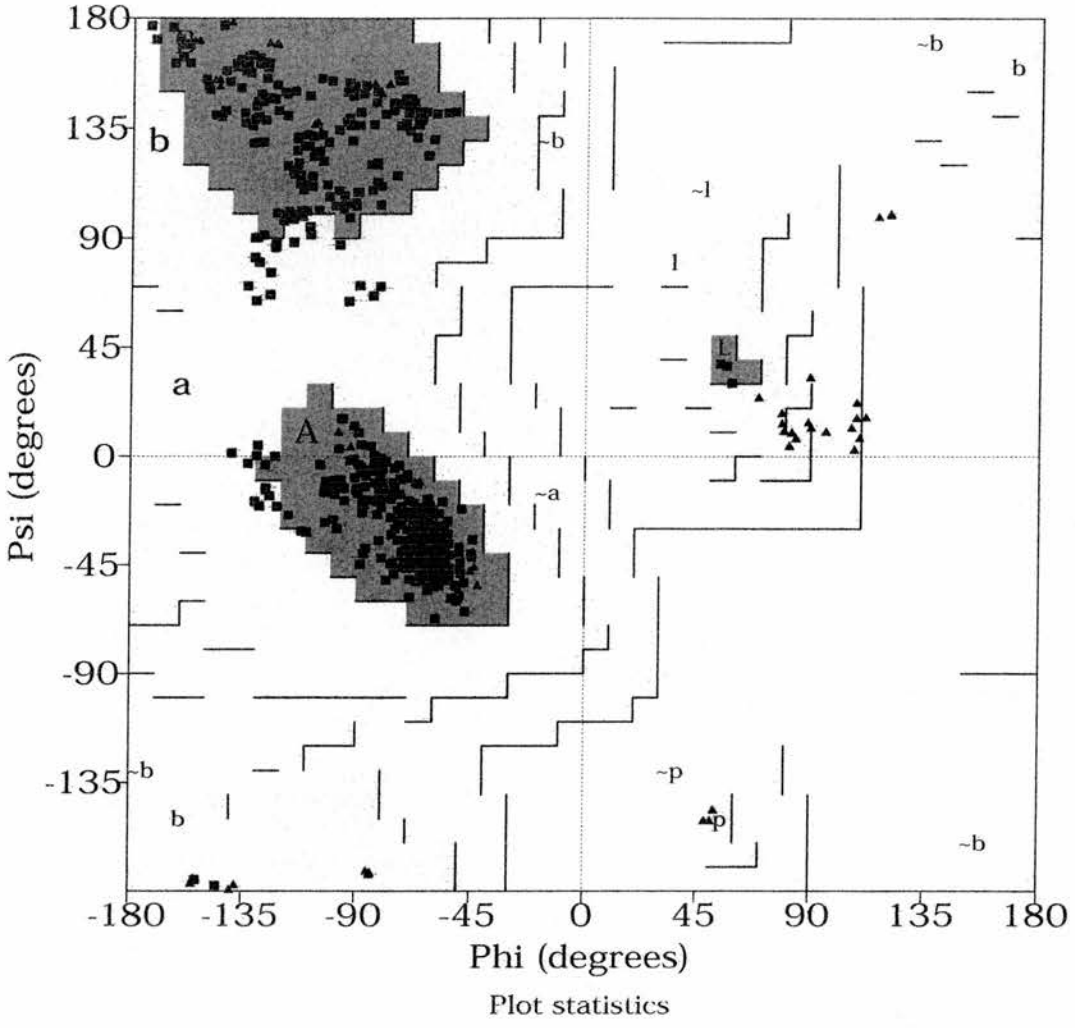


Figure 4.04. Ramachandran plot for the *E. coli* KDPG aldolase model

4.4 Results & Discussion

General Overview

KDPG aldolase from *E. coli* is a trimer, with each identical monomer being an α/β barrel. The trimer is shown in figure 4.05, a monomer is shown in figure 4.06. The sequence with the secondary structure elements aligned is shown in figure 4.06. These monomers are relatively simple α/β barrels, having an additional long helix at the N-terminus that caps the barrel. The helix at the C-terminus is slightly distorted and as such does not pack against the β sheets. Overall its topology is essentially identical to other TIM barrels and, naturally, *P. putida* KDPG aldolase.

E. coli and P. putida Structure Comparisons

As mentioned in Chapter Two, there is 45% identity between the enzyme from the two sources, and 63% similarity. Superposition of the two structures by sequence gave a residue by residue r.m.s. deviation, for all 213 C-α-atoms, of 5.1Å. Superposition of topologically equivalent regions, disregarding sequence, gave an r.m.s. deviation of 1.7Å for 177 C-α-atoms. The catalytic lysine of *P. putida* is located in an entirely different part of the structure compared to the catalytic lysine in *E. coli*. If the *P. putida* structure were correct it would be completely different to all of the other aldolase enzymes studied. The enzyme would also operate under a unique mechanism. Neither of these is likely. Given the high degree of sequence identity between the enzymes, the well-characterised aldolase fold, and topological similarity between the C-α backbones, the differences between them are errors in the *P. putida* structure rather than any real structural differences between the enzymes.

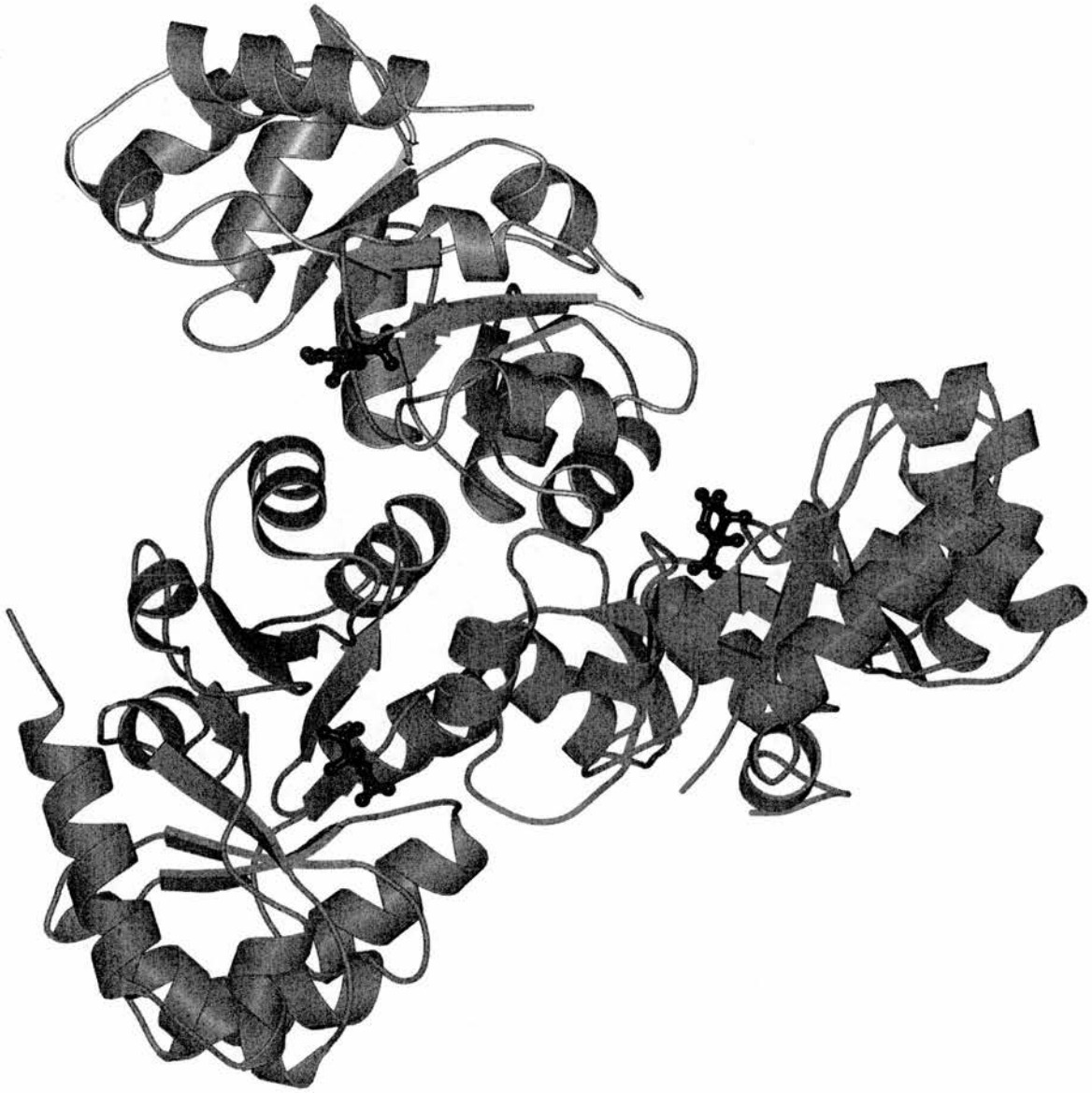


Figure 4.05. The trimeric arrangement of KDPG aldolase. The citrate anion bound in the active site is shown in black

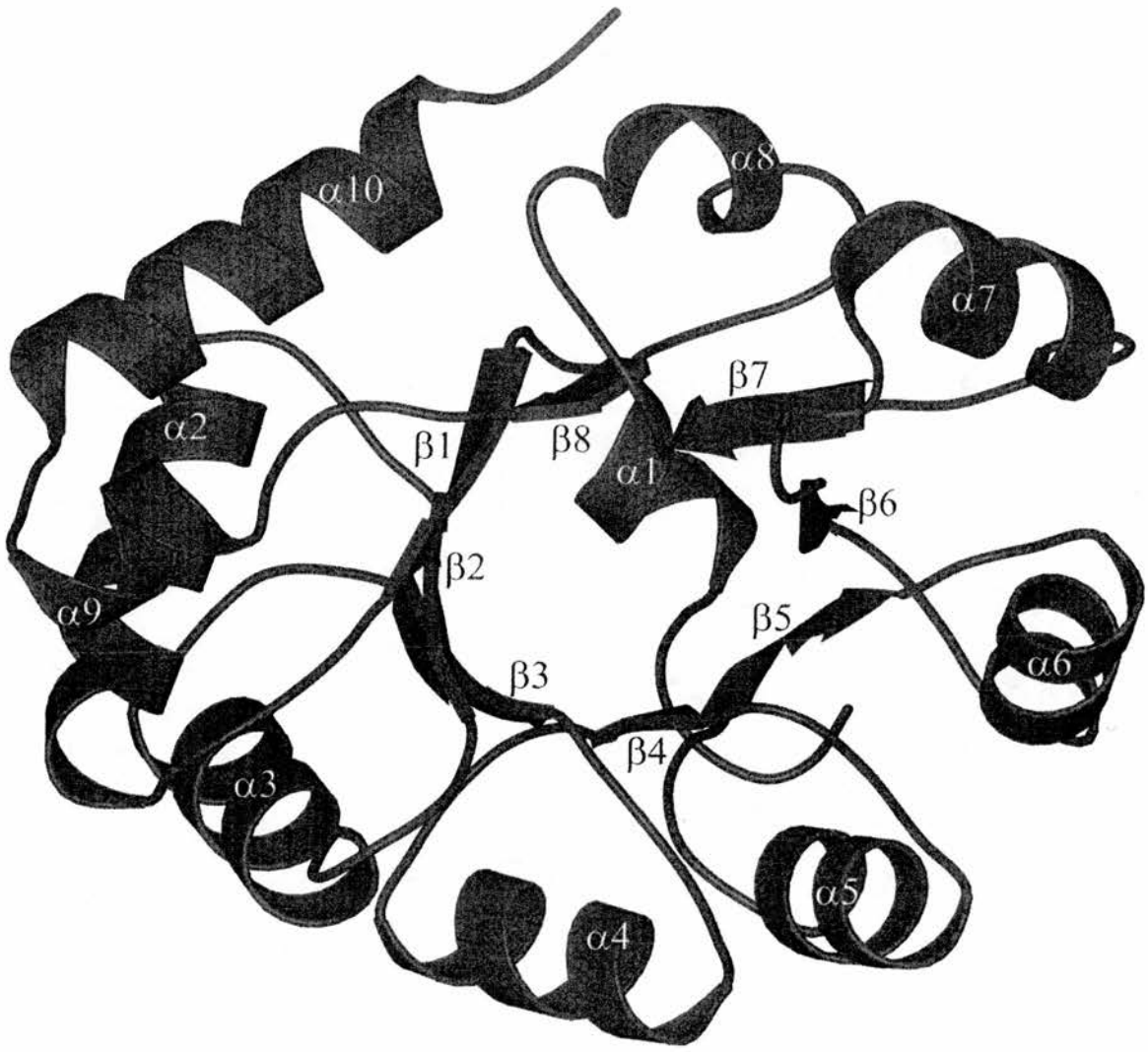


Figure 4.06. A monomer of KDPG aldolase, with the α helices and β sheets labelled

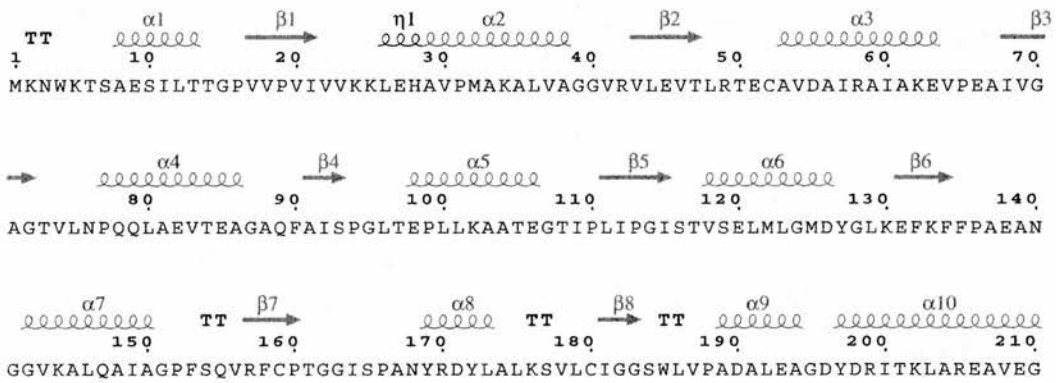


Figure 4.07. The secondary structure elements aligned with the KDPG aldolase sequence

Molecular Replacement Attempts – Post-Mortem

As described earlier the actual differences found between the enzyme from the two sources are due to errors in the *P. putida* structure rather than any true differences. From superpositioning of the C-α backbones of the two structures, figure 4.08, it can be seen that whilst the overall fold is relatively similar the actual arrangement of the two enzymes is completely different.

Trimer

The question 'why is KPDG aldolase from both *E. coli* and *P. putida* trimeric when all other aldolase enzymes are either dimeric or tetrameric ?' is posed. The final model of the *E. coli* enzyme was analysed using the Protein-Protein Interaction Server at University College London. This server analyses the structure and reports on protein interface interactions. A summary of these may be found in table 4.07. The axis of the three barrels lies in a plane that is at 90° to the non-crystallographic three-fold axis. The trimer buries approximately 3300Å^2 (1100Å^2 per monomer) of surface area and is composed of 57 van der Waals interactions. This, plus the 9 H-bonds, is low compared with other multimer interactions. As the same trimeric fold has been identified for both the *E. coli* and *P. putida* enzymes it does indicate the protein is trimeric in solution. Previous analysis by dynamic light scattering showed, from the estimated molecular weight, that the protein was likely to be trimeric in solution. To further investigate this, native gels were run to determine the molecular weight of the protein in solution. These gels proved inconclusive.

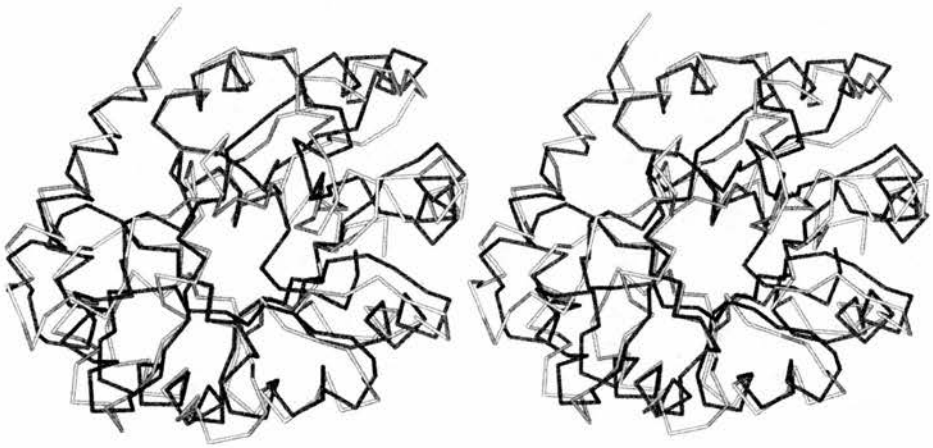


Figure 4.08. Stereo superpositioning of the C-α backbones of KDPG aldolase from *E. coli* and *P. putida*

Protein Interface Parameter	Result for *E. coli* KDPG aldolase
Interface accessible surface area	591
% Interface accessible surface area	6.03
% Polar Atoms in interface	30
% Non-Polar atoms in interface	71
Hydrogen bonds	3
Bridging water molecules	0

Table 4.07. Statistical analysis of monomer-monomer interactions in *E. coli* KDPG aldolase.

Active Site

The catalytic residue, Lys133, is located inside the barrel on β6 at the C-terminal, or open end, of the barrel. This is identical to all except two Class I Schiff base forming aldolase enzymes. Lys133 was confirmed as being the catalytic residue by MALDI-TOF. The KDPG aldolase-pyruvate Schiff-base complex was trapped by addition of cyanoborohydride and trypsin digested. Figure 4.09 shows the spectra for the pure native and pure native plus pyruvate complex, indicating only one pyruvate is bound in the sample. Figure 4.10 shows the trypsin digested spectra, the absence of peak 1613 in the bottom spectra is conclusive evidence Lys133 is the Schiff base forming residue.

In transaldolase the catalytic lysine is located on β4 (Jia *et al.*, 1996). However there is some very persuasive evidence to suggest a circular permutation of the gene has occurred and thus, in a structural sense, the lysine is found on the same strand as all other aldolases; β6. The second example is 5-aminolaevulinic acid dehydratase; here the catalytic lysine is shifted relative to the secondary structural elements (Erskine *et al.*, 1997). However, this enzyme is classified as a hybrid aldolase and its mechanism is not known.

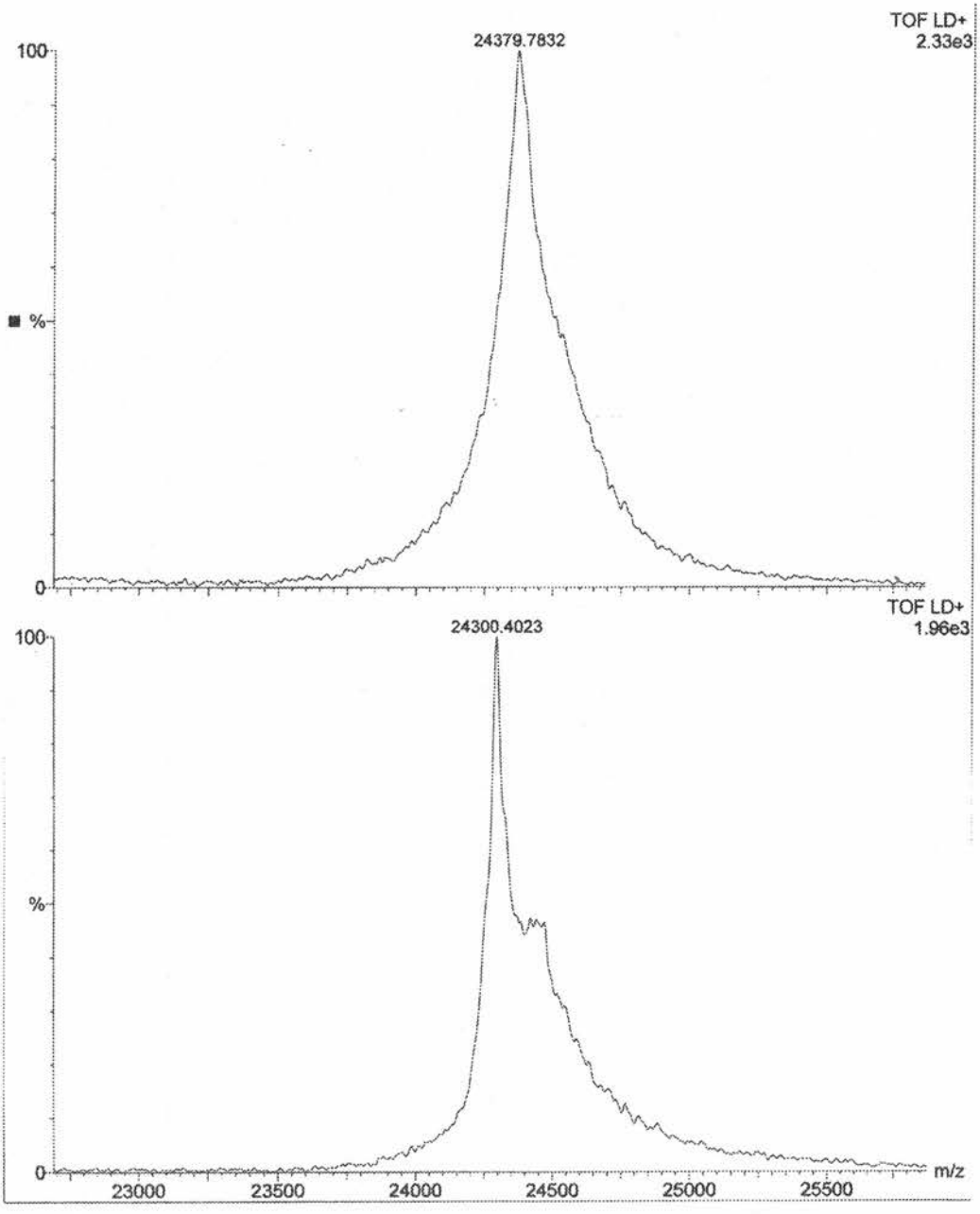


Figure 4.09 Top: Native aldolase + pyruvate. Bottom: Native aldolase only.

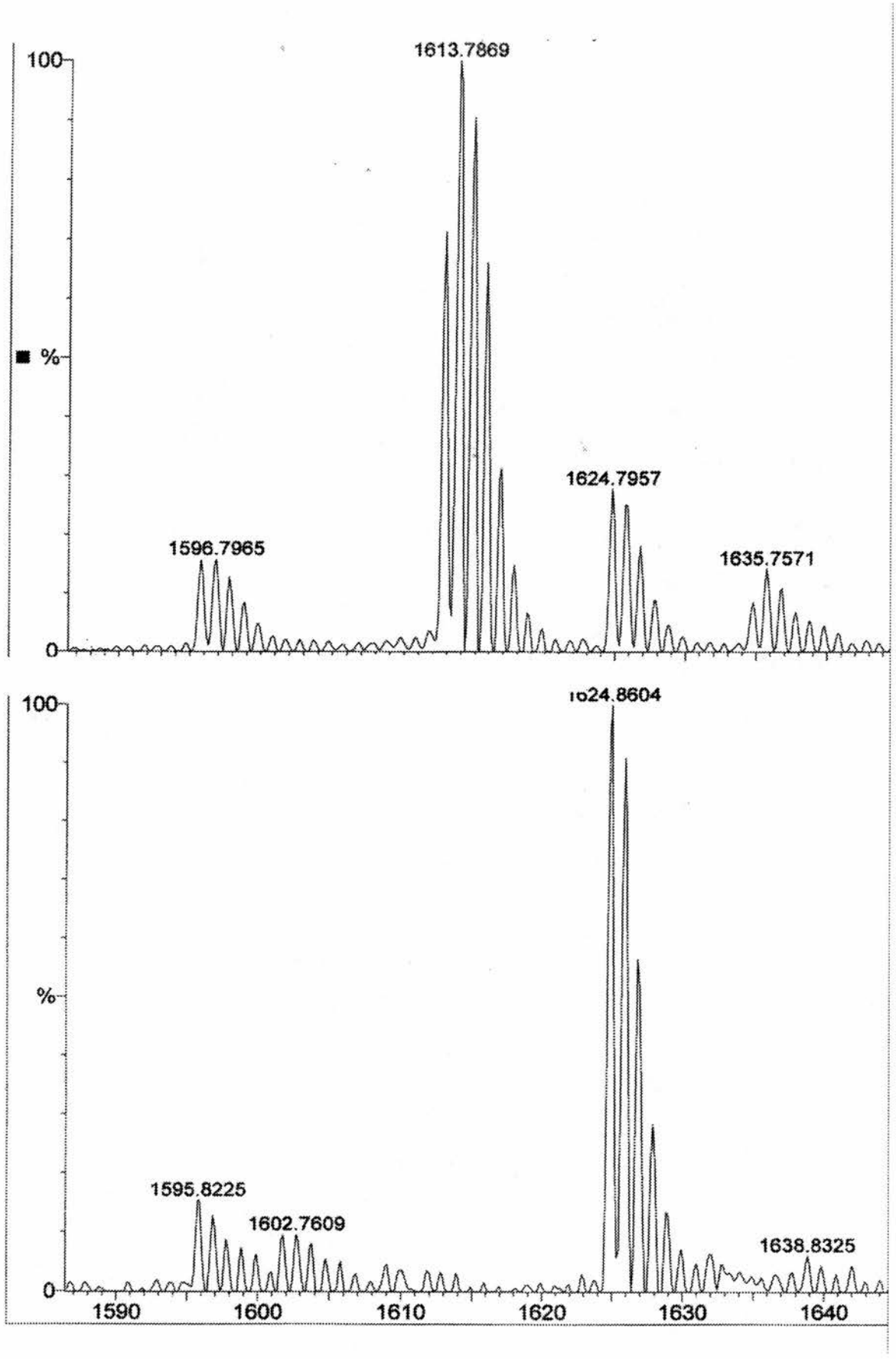


Figure 4.10 Top: Digested Native only. Bottom: Digested Native + pyruvate.

The active site is lined with residues Val20, Glu45, Arg49, Thr73, Phe135 and Thr161. These residues are exclusively located in highly conserved motifs of the KDPG aldolase family. The citrate anion mentioned during refinement makes Hydrogen bonds to Glu45, Arg 49, Thr73, Lys133 and Thr161. There is a large hydrophobic pocket adjacent to the active site formed by the inner walls of the barrel, with a diameter of approximately 7Å. This pocket is lined with residues Val18, Val20, Phe90, Ile92, Ile112 and Cys159. A summary to these residues and the interactions they are involved in is shown in table 4.08. The arrangement of the active site is shown in figure 4.11.

Residue	Active Site	Citrate Binding	Hydrophobic Pocket
Val 18	No	No	Yes
Val 20	Yes	No	Yes
Glu 45	Yes	Yes	No
Arg 49	Yes	Yes	No
Thr 73	Yes	Yes	No
Phe 90	No	No	Yes
Ile 92	No	No	Yes
Ile 112	No	No	Yes
Lys 133	Yes	Yes	No
Phe 135	Yes	No	No
Cys 159	No	No	Yes
Thr 161	Yes	Yes	No

Table 4.08. Residues involved in the active site of *E. coli* KDPG aldolase

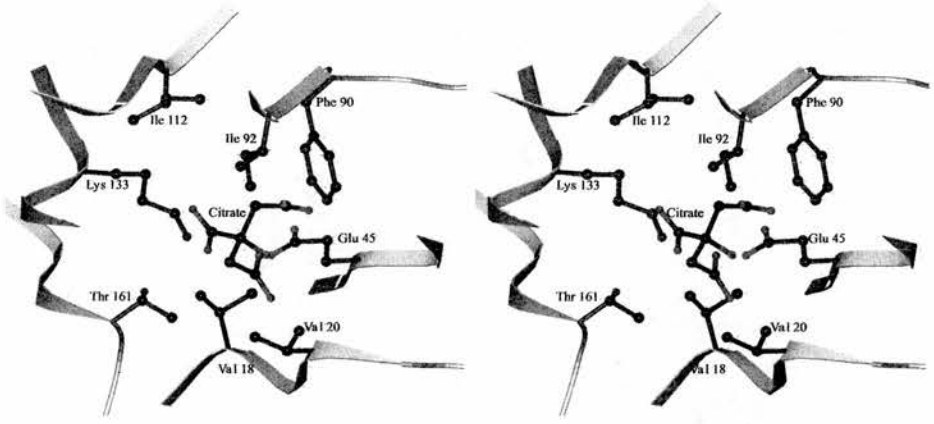


Figure 4.11. The active site of KDPG aldolase with bound citrate

The hydrogen-bonding network from monomer B is shown in figure 4.12. The citrate anion O6 binds to Lys133; O2 and O3 bind to Thr73. O4 is involved in hydrogen bonding via water molecules and Arg49 is bonded to O4. Bonding to Gly45, the catalytic base, is via a water molecule.

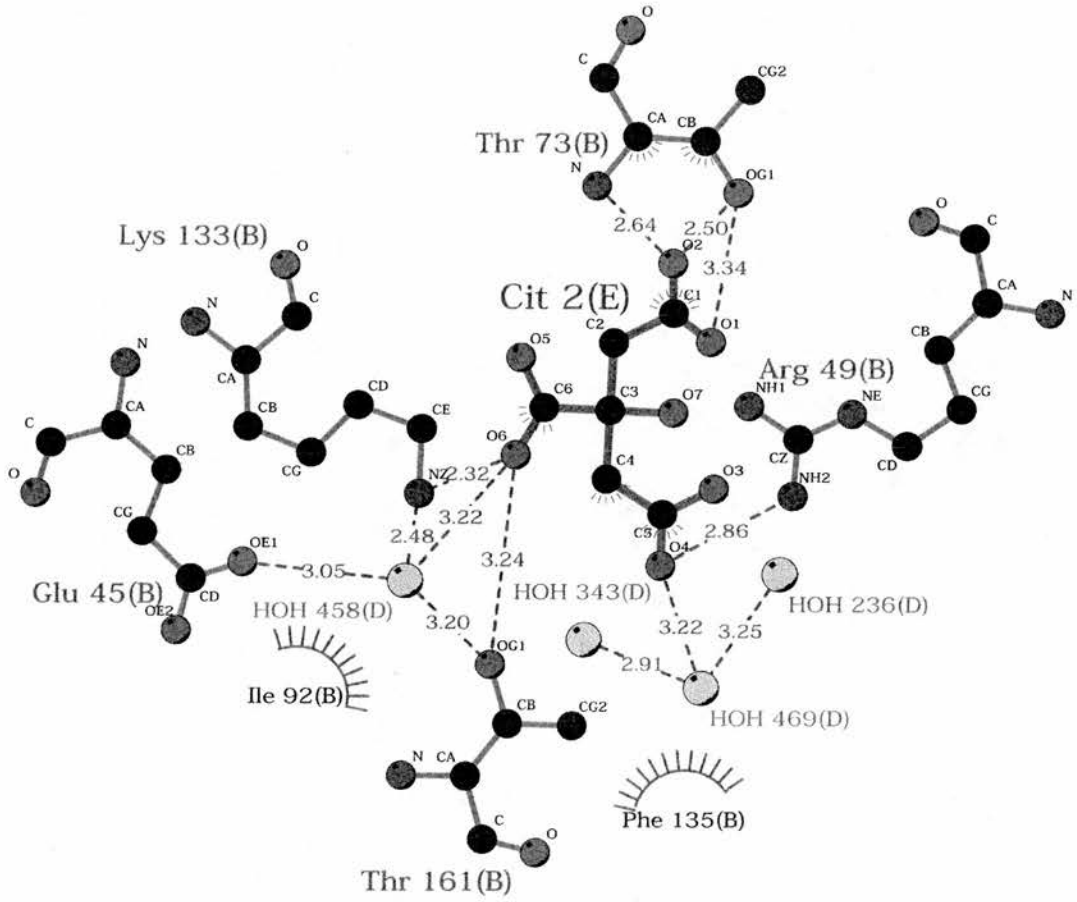


Figure 4.12. The Hydrogen bonding network in monomer B

Enzyme Mechanism

A unified mechanism for Class I aldolases proposed from studies on transaldolase and FBP aldolase has been proposed. This mechanism proposes a principle catalytic base and a secondary catalytic acid in addition to a Schiff base forming lysine.

In *E. coli* KDPG aldolase Glu45 aligns with aspartic acids in both transaldolase (Asp17) and FBP (Asp33). These aspartic acids are the principle catalytic bases. Therefore it seems evident that Glu45 is acting as the principle catalytic base in KDPG aldolase. Analysis of the secondary catalytic acid is not so straightforward.

A glutamic acid, Glu96 in transaldolase and Glu187 in FBP aldolase, acts as a catalytic acid promoting the formation of the Schiff base complex. However, it is now known that a tyrosine is found at this position in N-acetylneuraminic acid lyase and the mechanism would have to be adjusted accordingly (Izard *et al.*, 1994). More importantly in all KDPG aldolases this residue is a glycine, Gly114, which cannot function as an acid.

To fully understand the mechanism, further investigations will have to be made. These will include the determination of the crystal structure of the trapped Schiff-base complex and site directed mutagenesis of residues thought to be involved.

PDB Accession Code

KDPG aldolase has been deposited in the Protein Data Bank. Its accession code is 1FQ0.

Chapter Five

Structure determination of the KDPG aldolase double mutant K133Q/T161K

5.1 Abstract

A mutant of KDPG aldolase, T161K, was identified by directed evolution as having broader substrate specificity than the native enzyme. Given the structure of aldolase enzymes it was thought this new lysine could be functioning as the Schiff base forming residue. This was confirmed by constructing the double mutant K133Q/T161K by site directed mutagenesis; this double mutant retained enzymatic activity. This mutant was cloned and expressed into the pET-28-b(+) vector. It was purified in an identical manner to the native KDPG aldolase and crystals obtained. The crystals were of much lower quality than their native counterparts. Diffraction from these crystals was found to be extremely variable; most did not diffract past 5Å even with synchrotron radiation. A 3.2Å data set was obtained on BM14 at ESRF Grenoble. This was solved in AMoRe using the refined 2.17Å native E. coli KDPG aldolase model. A citrate anion was found bound in the active site. Very limited refinement gave an R_{factor} of 24.0% and an R_{free} of 29.6%. The overall fold of the mutant protein was found to be identical to the native. At this resolution analysis of the active site was limited. A 2.7Å data set was obtained on ID14-1 at ESRF Grenoble. This was solved in AMoRe using the 2.17Å native E. coli KDPG aldolase data. A citrate anion was found bound in the active site. An R_{factor} of 23.8% and R_{free} of 27.6% were obtained. With this higher resolution data a much more complete analysis of the active could be made.

This is the first example of an aldolase Schiff base lysine being moved from strand β6 to strand β7. The structure explains how activity is retained and how the substrate specificity has been broadened.

5.1 Introduction

The specificity of enzymes is one of their features those who utilise them would like to be able to alter. This ability to manipulate enzymes has been theoretically possible for many years. To do this active site residues, identified through either site directed mutagenesis or crystallographic studies, would be mutated to enable them to bind new substrates and thus catalyse the desired reaction. However this method has proven fallible, primarily because the complex interactions between altering residues and the impact this has on protein function are poorly understood. Thus over the last two to three years the emphasis has moved from site-directed mutagenesis to directed evolution as a means of developing new enzymes with altered substrate specificity (Spiller *et al.*, 1999).

The current methods of directed evolution have been previously described (Chapter One). These methods have become increasingly common; for example this approach has been used to alter enzyme substrate specificity and thermophilic enzyme properties (Lebbink *et al.*, 2000; Merz *et al.*, 2000). A landmark in this field was the work of Fersht's group on directed evolution using α/β barrel proteins (Altamirano *et al.*, 2000).

In this study the enzyme under investigation had its substrate binding residues primarily located within the loop regions while the catalytic residues were predominantly located within the barrel. This partition of specificity and catalysis between structurally distinct regions is a convenient device for evolution to generate diversity in a combinatorial manner. It was this structural feature which raised the possibility of mimicking

evolution *in vitro* either by moulding a new catalytic function into an existing binding site, or a new binding site into an existing catalytic activity.

The monomeric α/β barrel protein indole-3-glycerol-phosphate synthase (IGPS) was used as the scaffold, and its activity switched to that of phosphoribosyl-anthranilate isomerase (PRAI). The two enzymes form two covalently linked domains of a bi-functional enzyme in *E. coli* which catalyse two consecutive steps in the tryptophan biosynthesis pathway. The catalytic residues were conserved in this approach, however the loops, which control substrate specificity, were allowed to vary in length and sequence. That this procedure was successful confirms α/β barrels as being particularly suited to manipulation by directed evolution strategies to evolve new activities.

5.3 Experimental

Directed Evolution

Directed evolution experiments were undertaken at Duke University, USA. No restrictions were placed on which residues were allowed to undergo mutation. Screening and selection of the mutant library, based on increased or altered substrate specificity, identified one mutant with broadened substrate specificity; T161K. As the native KDPG aldolase shows a symmetrical disposition between T161 and K133 about a residue implicated in the reaction mechanism, Glu45, it was though this new lysine could act as the catalytic residue and also be responsible for the altered substrate specificity. To address whether this new lysine was acting as the Schiff base forming lysine the double mutant K133Q/T161K was constructed by site-directed mutagenesis to remove the catalytic lysine. This double mutant showed the same broad substrate specificity as the single mutant T161K. It was decided to investigate the double mutant crystallographically to determine a structural basis for the transferral of the Schiff base forming lysine from strand β6 to β7 and the resultant increased substrate specificity profile.

The double mutant was cloned into the pET28-b(+) vector and on receiving this clone in St Andrews it was transformed into the BL21(DE3)pLysS cell line.

Enzyme Activity & Specificity

The double mutant, K133Q/T161K, KDPG aldolase was found by the Toone group, Duke University, to show a markedly altered substrate specificity with respect to the native enzyme. More specifically the mutant enzyme showed enhanced activity against

pyridine carboxyaldehyde, benzaldehyde and α-ketobutyrate. To assess this, and the level of asymmetric induction provided by the mutant, a preparative scale addition of pyruvate to 2-pyridine carboxyaldehyde was carried out. By comparison to authentic material of known absolute configuration the mutant yields *S*-4-hydroxy-α-ketobutyrate with an optical purity of 32%. This enantiofacial selectivity is remarkable given the significant modification of the enzyme active site.

Protein Purification

A single colony was isolated from an Luria Agar plus kanamycin plate and used to inoculate an overnight culture of 10ml LB plus 100μg ml^{-1} kanamycin and 50μg ml^{-1} chloramphenicol, incubated at 310K, 175 rpm shake. This culture was then used to inoculate 2L of TB plus 100μg ml^{-1} kanamycin and 50μg ml^{-1} chloramphenicol and incubated at 310K, 175 rpm shake. The growth was monitored until an OD_{600} of 0.6 was reached, IPTG was then added to a final concentration of 0.4mM and incubation continued for a further 3 hours. The cells were harvested by centrifugation, 8000 g, 15 mins, 277K and the resultant pellet frozen at 253K. The pellet was defrosted on ice and resuspended in 20mM Tris-base, pH 8.0, 1mM EDTA, 1mM DTT at a concentration of 10ml L^{-1}. Following incubation at room temperature for 30 mins the viscosity of the solution was reduced by the addition of 20μg ml^{-1} DNAse I. To ensure complete cell lysis occurred the cells were sonicated on ice, 6 cycles of 30 s with 30 s rest. The cell debris was removed by centrifugation, 18000g, 15 mins, 277K. The resultant supernatant was brought to 20% $(NH_4)_2SO_4$ at 277K and centrifuged to pellet insoluble non-induced proteins, 18000 g, 15 mins, 277K. This supernatant was dialysed against 20mM Tris-base, pH 8.0 to remove all salt prior to the initial purification step and filtered through a 0.2μm membrane.

This protein sample was applied to a POROS-HQ HPLC column, BioCad 700E, and eluted against an increasing salt gradient of 0-1000mM NaCl, 20mM Tris-base, pH 8.0. Coomassie stained SDS-PAGE analysis of the fractions showed a protein at 23kDa to be eluted at 100-120mM NaCl; this was identified as the mutant KDPG aldolase.

The mutant KDPG aldolase sample was digested with thrombin, to cleave the His-tag from the protein, at a concentration of 10units mg^{-1} fusion protein, for 16 hours at 295K. It was brought to 30% $(NH_4)_2SO_4$ saturation at 277K. This was then applied to a POROS high density phenyl HPLC column, BioCad 700E, previously equilibrated in 30% $(NH_4)_2SO_4$, 20mM Na_2HPO_4, pH 8.0. The protein was eluted in a decreasing salt gradient, 30 to 0% $(NH_4)_2SO_4$. Silver stained SDS-PAGE analysis of the fractions showed KDPG aldolase to be eluted at 10% $(NH_4)_2SO_4$ and be present as a single band. The fractions corresponding to this peak were combined and concentrated under nitrogen pressure in an Amicon ultrafiltration unit at 277K. N-terminal sequencing and mass spectroscopy confirmed the protein's identity.

Crystallisation

The mutant protein was used in crystallisation experiments using the previously optimised conditions of 20% PEG 6K, 0.075M citric acid, pH 4.0, 30% sucrose at a protein concentration of 3mg ml^{-1}. Experiments were set up in both sitting and hanging drop trays. Protein crystals were obtained which were of much lower quality than their native and Se-Met counterparts; they were smaller and much more fragile.

The initial crystallisation screens, Hampton I & II, were performed with the mutant protein to obtain better crystals. The Hampton additive screen was also repeated with the mutant protein. Extensive searching of conditions found the best crystals were obtained with the initial conditions used.

Data Collection

Copper Rotating Anode

Crystals were soaked for 30 s in 15% glycerol prior to data collection at 130K. With an impending trip to ESRF Grenoble, crystals were initially analysed to ensure they diffracted. A selection of the crystals were found to diffract beyond 4Å thus it was decided to collect data at the ESRF Grenoble rather than in-house with the aim of recording much higher resolution data.

Grenoble BM14

After screening between 40 and 50 crystals, none of which diffracted past 5Å, data were collected from a single crystal on station BM14 the ESRF synchrotron at Grenoble using an ADSC Quantum-4 CCD detector. The crystal was soaked for 30 s in the mother liquor plus 15% glycerol prior to data collection at 100K. Data were recorded as 180 images with an exposure time of 20 s, as 0.5° non-overlapping images. Data were indexed and integrated with *MOSFLM*, and merged with *SCALA* (Leslie, 1992) to a resolution of 3.2Å. The data gave an R_{merge} of 12.6%, with unit cell parameters of a = 70.4Å, b = 68.6Å, c = 127.2Å with $\alpha = \beta = \gamma = 90°$. A summary of the quality of the data may be found in table 5.01.

Resolution (Å)	No. Reflections	Completeness (%)	R_{merge} (%)	Redundancy
30 – 7.85	776	97.9	2.9	3.3
7.85 – 6.25	735	100	6.1	3.5
6.25 – 5.46	723	100	10.1	3.6
5.46 – 4.97	717	100	9.6	3.6
4.97 – 4.61	719	100	10.7	3.7
4.61 – 4.31	704	100	8.9	3.7
4.31 – 4.12	689	100	9.7	3.8
4.12 – 3.94	716	100	12.4	3.7
3.94 – 3.79	701	100	13.3	3.7
3.79 – 3.66	690	100	18.5	3.8
3.66 – 3.55	702	100	22.1	3.7
3.55 – 3.45	695	100	29.0	3.8
3.45 – 3.36	698	100	30.7	3.8
3.36 – 3.27	690	100	36.4	3.8
3.27 – 3.20	698	100	47.7	3.8
30 – 3.2	10643	99.8	12.6	3.6

Table 5.01. 3.2Å Mutant KDPG aldolase data statistics

Molecular Replacement at 3.2Å

The structure of the mutant KDPG aldolase was determined using molecular replacement as implemented in the CCP4 program *AMoRe* with data from 15 to 4Å (CCP4, 1994; Navaza & Saludjian, 1997). The trimeric native *E. coli* KPDG aldolase refined structure was used as the search model. The solutions from each stage of the *AMoRe* process are shown in tables 5.02 to 5.04.

Solution	α	β	γ	c.c*	σ
1	128.06	47.49	281.14	26.9	1.7
2	127.68	46.10	41.12	26.0	1.8
3	125.80	47.15	162.71	25.2	18

Table 5.02. Solutions from cross-rotation function

Solution	α	β	γ	x	y	z	c.c*	R_{factor}	σ
1	128.06	47.79	281.14	0.1583	0.9997	0.1525	53.0	43.8	2.3
2	127.86	46.10	41.12	0.1587	0.0017	0.1525	56.2	43.1	2.3
3	125.80	47.15	162.71	0.1580	0.9997	0.1527	49.0	46.2	2.3

Table 5.03. Solutions from translation function

Solution	α	β	γ	x	y	z	c.c*	R_{factor}
1	128.1	46.14	253.82	0.1558	-0.0018	0.1505	60.1	40.6

*c.c is the correlation coefficient

Table 5.04. Solution from fitting function

Refinement at 3.2Å

Prior to commencing refinement 5% of the data were removed from all further calculations to give a free-R_{factor} from which the quality of refinement could be measured statistically. The same 5% set of reflections was chosen for this data as for the native 2.1Å refinement. The refinement program CNS (Brunger *et al.*, 1998) was used to refine the structure, however due to the low resolution of the data only one round of positional refinement, 300 cycles, and one round of temperature factor refinement, 30

cycles, were ran with tight NCS restraints. A complete analysis of the final model may be found in table 5.05.

Statistic	Mutant KDPG aldolase at 3.2Å
R_{factor}●	24.0%
R_{free}	29.6%
R_{factor} (3.27-3.2Å)	30.2%
R_{free} (3.27-3.2Å)	37.5%
Average B-factor■	36Å
RMSD* bonds	0.0083Å
RMSD* angles	1.404°
B-factor Deviation bonds■	1.16Å^2
B-factor Deviation angles■	2.07Å^2
Residues in most favoured region of Ramachandran	84.9%
Number of non-hydrogen atoms	4743
Number of atoms set to zero occupancy	0

*RMSD calculated via *LSQMAN* ●The difference between the absolute value of the observed structure factor amplitudes from the data & those calculated from the model, for all reflections, divided by the sum of the observed structure factor amplitudes ■B-factor deviation for bonded atoms

Table 5.05. Statistical analysis of the 3.2Å mutant KDPG aldolase model

Additional visits to Grenoble using beamlines BM14, ID14-1 and ID14-2, typically screening over 50 crystals each time, yielded none that diffracted past 3.2Å.

Grenoble ID14-1

After screening between 10 and 15 crystals data were collected from a single crystal on station ID14-1 at the ESRF Grenoble synchrotron using an ADSC Quantum-4 CCD detector. The crystal was soaked for 30 s in the mother liquor plus 15% glycerol prior to data collection at 100K. Data were recorded as 180 images with 20 s exposure, as 0.5°

non-overlapping images. Data were indexed and integrated with *MOSFLM*, and merged with *SCALA* (Leslie, 1992) to a resolution of 2.7Å. The data gave an R_{merge} of 8.7%, with unit cell parameters a = 54.87Å, b = 84.5Å, c = 135.0Å with $\alpha = \beta = \gamma = 90°$. A summary of the quality of the data may be found in table 5.06.

Resolution (Å)	No. Reflections	Completeness (%)	R_{merge} (%)	Redundancy
67 – 8.54	1739	97.9	6.5	2.8
8.54 – 6.04	3270	99.1	6.3	3.1
6.04 – 4.93	4327	99.4	6.7	3.2
4.93 – 4.27	5188	99.6	6.4	3.3
4.27 – 3.82	5802	99.7	7.0	3.3
3.82 – 3.49	6446	99.8	8.4	3.3
3.49 – 3.23	6959	99.8	10.7	3.3
3.23 – 3.02	7621	99.8	13.2	3.3
3.02 – 2.85	8086	99.9	18.8	3.4
2.85 – 2.70	8553	99.8	25.6	3.4
67-2.7	57991	99.9	8.7	3.3

Table 5.06. 2.7Å mutant KDPG aldolase data statistics

Molecular Replacement at 2.7Å

The structure of the mutant KDPG aldolase was determined using molecular replacement as implemented in the CCP4 program *AMoRe* with data from 15 to 3Å (CCP4, 1994; Navaza & Saludjian, 1997). The trimeric native *E. coli* KPDG aldolase was used as the search model. All solutions from each stage of the *AMoRe* process are shown in tables 5.07 to 5.09. All solutions are related by 120°.

Solution	α	β	γ	c.c*	σ
1	173.34	59.41	111.03	27.7	4.7
2	174.92	58.53	230.24	27.4	4.7
3	173.63	57.84	351.31	26.4	4.7

Table 5.07. Solutions from cross-rotation function.

Solution	α	β	γ	x	y	z	c.c*	R_{factor}	σ
1	172.95	60.00	111.22	0.2441	0.3035	0.1635	65.8	42.2	5.4
2	173.80	59.10	230.84	0.2441	0.3046	0.1634	64.6	43.1	5.4
3	173.07	58.86	351.50	0.2445	0.3028	0.1626	61.4	45.3	5.4

Table 5.08. Solutions from translation function.

Solution	α	β	γ	x	y	z	c.c*	R_{factor}
1	173.34	59.41	111.03	0.2472	0.3039	0.1632	53.9	45.8
2	174.92	58.53	230.24	0.2461	0.3037	0.1634	51.9	47.0
3	173.63	57.84	351.31	0.2463	0.3032	0.1624	47.9	49.2

*c.c is the correlation coefficient

Table 5.09. Solution from fitting function

Refinement at 2.7Å

Prior to commencing refinement 5% of the data were removed from all further calculations to give a free-R_{factor} from which the quality of refinement could be measured statistically. The same 5% set of reflections was chosen for this data as for the native 2.1Å refinement. The refinement program CNS (Brunger *et al.*, 1998) was used

to refine the structure; one round of positional refinement consisted of 300 cycles, and one round of temperature factor refinement consisted of 30 cycles.

The model was subjected to refinement and manipulated manually in *O* (Jones *et al.*, 1991), subjected to further refinement, once more viewed in *O* and had water atoms added in CNS. Water-pick was used to select water atoms which satisfied the conditions described in Chapter 4. Once the waters were added they were subjected to further rounds of refinement and examined manually in *O*. In an identical manner to the native KDPG aldolase a citrate anion was found in the active site. A complete analysis of the final model may be found in table 5.10.

Statistic	Mutant KDPG aldolase at 2.7Å
R_{factor}●	23.8%
R_{free}	27.6%
R_{factor} (2.76-2.70Å)	27.4%
R_{free} (2.76-2.70Å)	35.3%
Average B-factor■	35Å
RMSD* bonds	0.0063Å
RMSD* angles	1.32°
B-factor Deviation bonds■	1.19Å^2
B-factor Deviation angles■	1.98Å^2
Residues in most favoured region of Ramachandran	91.3%
Number of non-hydrogen atoms	4881
Number of atoms set to zero occupancy	330

*RMSD calculated via *LSQMAN* ●The difference between the absolute value of the observed structure factor amplitudes from the data & those calculated from the model, for all reflections, divided by the sum of the observed structure factor amplitudes ■B-factor deviation for bonded atoms

Table 5.10. Statistical analysis of 2.7Å mutant KDPG aldolase model

PROCHECK was run as part of CCP4i (CCP4, 1994). The statistics from the Ramachandran plot (Ramakrishnan & Ramachandran, 1965) are in table 5.11. The actual plot is shown in figure 5.01.

Statistic	No. of Residues	% of Residues
Residues in most favoured regions	585	91.5
Residues in additional allowed residues	28	8.5
Residues in generously allowed regions	0	0
Residues in disallowed regions	0	0

Table 5.11. Statistical analysis of the Ramachandran plot for the mutant KDPG aldolase

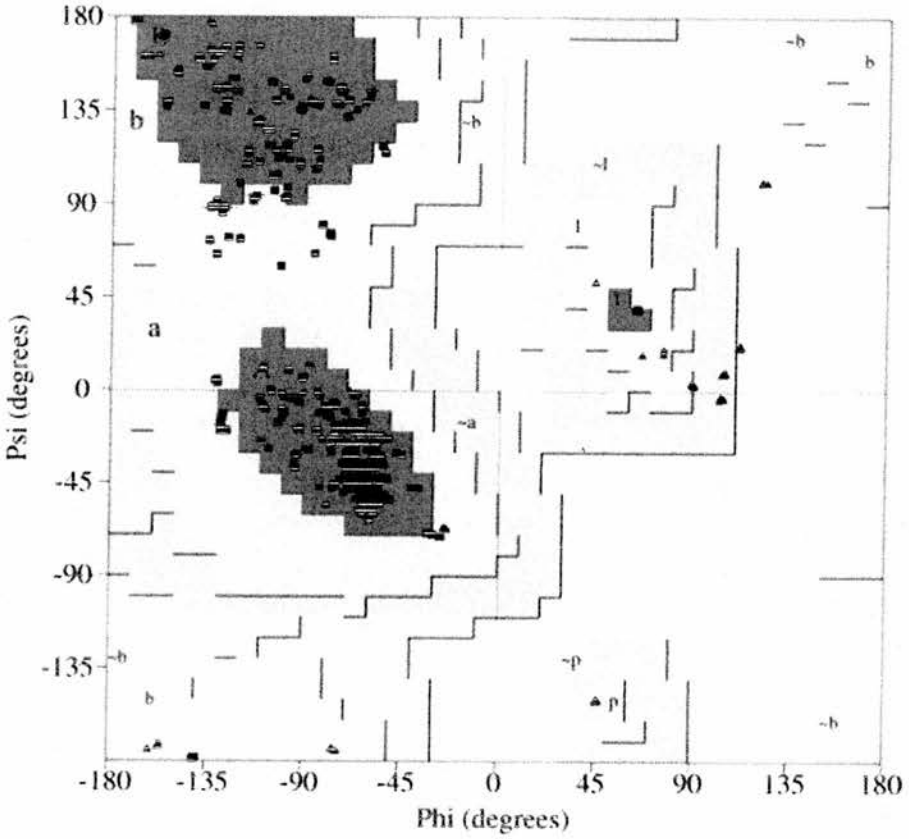


Figure 5.01. Ramachandran plot for mutant KDPG aldolase at 2.7Å

5.4 Results & Discussion

Purification & Crystallisation

The mutant protein was purified in an identical manner to the native protein. It also crystallised under identical crystallisation conditions. However, the crystals were of variable quality; their seemingly random range of diffraction was from 10Å to 2.7Å at the ESRF Grenoble synchrotron on beamlines BM14, ID14-1 and ID14-2. After initially solving and refining the mutant structure to 3.2Å data were obtained to 2.7Å and the structure refined to this higher resolution.

General Overview

The double mutant K133Q/T161K *E. coli* KDPG aldolase was found to have an overall topology identical to the native KDPG aldolase from *E. coli*. The average r.m.s. deviation for a C-α mutant monomer compared with a native monomer is 0.46Å. When variable loops are excluded the r.m.s. deviation is 0.34Å. A stereo superposition of native and mutant monomers is shown in figure 5.04.

Mass Spectroscopy

The catalytic residue, Lys161, is located inside the barrel on β7 at the C-terminal, or open end, of the barrel. It has been shown that this mutated lysine is acting as the Schiff-base forming residue by MALDI-TOF spectroscopy. The pyruvate-Schiff base complex was trapped with cyano-borohydride and subjected to trypsin digestion. The spectra obtained are shown in figure 5.02. For the mutant protein, the top spectra, the peak shown at 1481 indicates the pyruvate bound to a fragment that only contains the new lysine, Lys161. This peak is not present in the pure mutant spectra shown below. Figure 5.03 shows the undigested spectra for both the pure mutant, bottom, and mutant

plus pyruvate, top. This spectra shows only one pyruvate is bound; conclusive evidence that the new lysine is acting as the Schiff base residue.

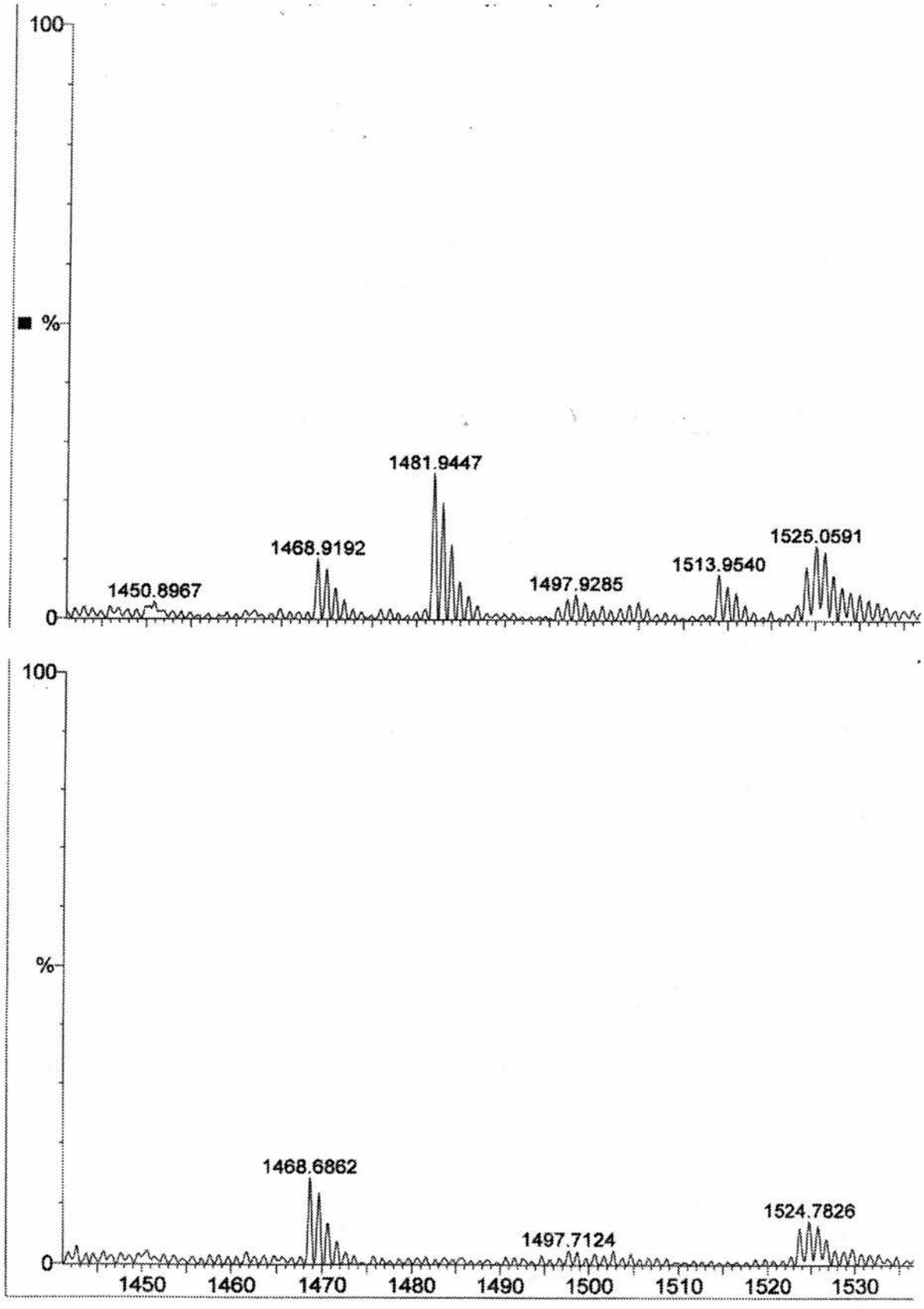


Figure 5.02. Schiff-base mutant enzyme complex top, mutant enzyme only bottom

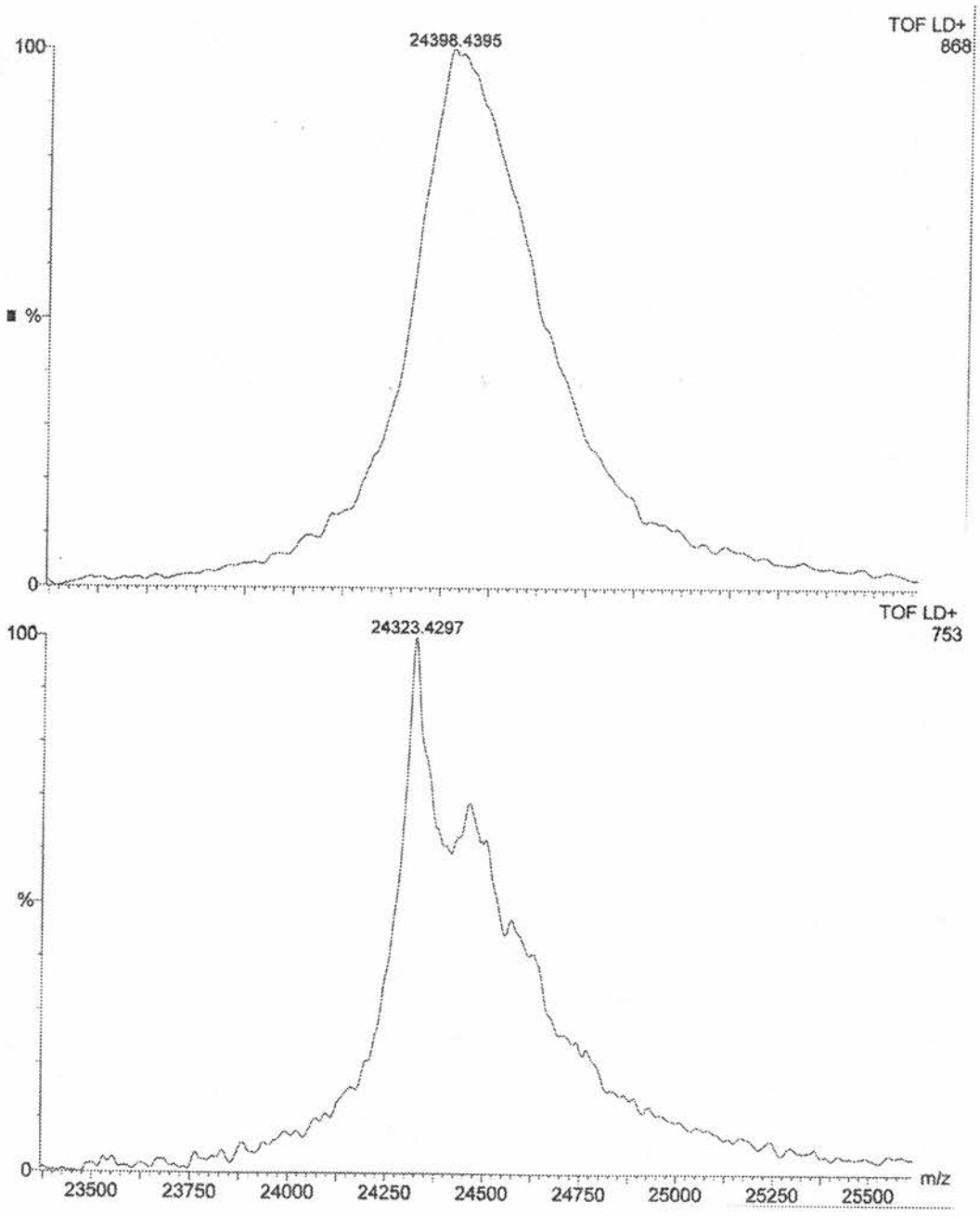


Figure 5.03. Undigested mutant plus pyruvate top, undigested pure mutant bottom

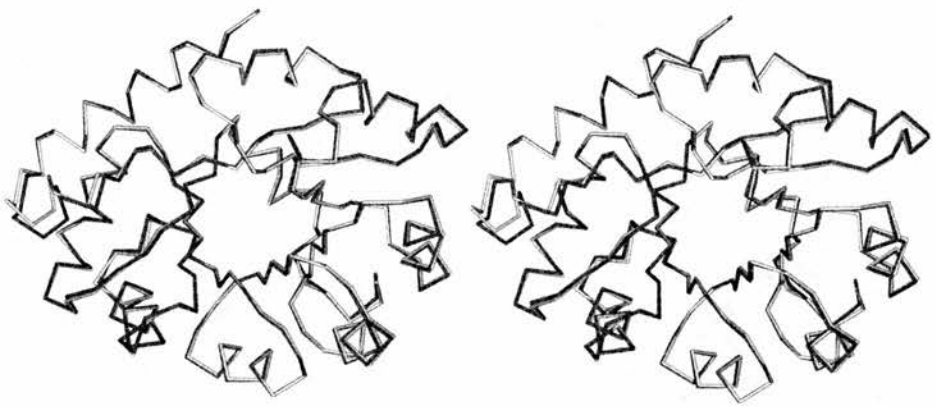


Figure 5.04. Stereo superposition of native (blue) and mutant (yellow) KDPG aldolase monomers

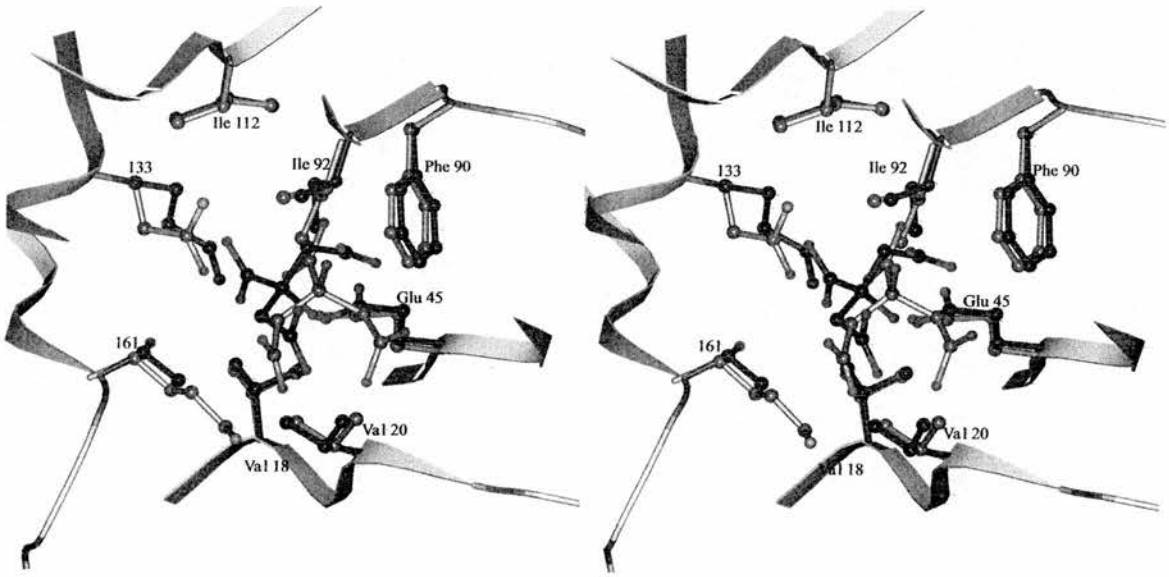


Figure 5.05. Stereo superposition of the active sites of the native and mutant KDPG aldolase. Residues are shown in ball and stick motif, with the protein main chain as a ribbon. C-atoms for the native are in dark grey; C-atoms for the mutant are light grey.

Active Site

In the mutant active site the new catalytic lysine, Lys161, is in a different configuration than the native, Lys133, however the citrate anion is shifted slightly to compensate for this. This is shown in figure 5.05. Originally it was thought the new lysine would orient itself so its tail was in the same position as the tail of the original lysine. However, from analysis of the 2.7Å data this is clearly not the case. All other non-mutated residues are located in essentially the same position in the mutant protein as the native protein. From the double mutant data it is thought the enzyme mechanism proposed for the native enzyme will still be accurate. A Hydrogen bonding network diagram for monomer A is shown in figure 5.06.

Models of the Schiff-base complexes for the native KDPG aldolase and the double mutant KDPG aldolase were constructed based on the structures of transaldolase and FBP substrate complexes. This model confirms the use of Glu45 as the principle catalytic base and indicates Arg49 may play a role in the binding of the terminal phosphate of the natural electrophilic substrate. These findings are consistent with the current model of aldolase function. This model also suggests the hydrophobic pocket identified in the native structure close to the active site lysine may accommodate larger and / or non-polar electrophilic substrates. An equivalent model of the mutant enzyme shows the electrophilic substrates to be shifted away from the hydrophobic pocket, thus effectively enlarging the size of the pocket and altering its electronic character. It is thought these changes result in the significant differences in both substrate specificity of the mutant enzyme and the level of asymmetric induction recorded during aldol addition when it is compared with the native.

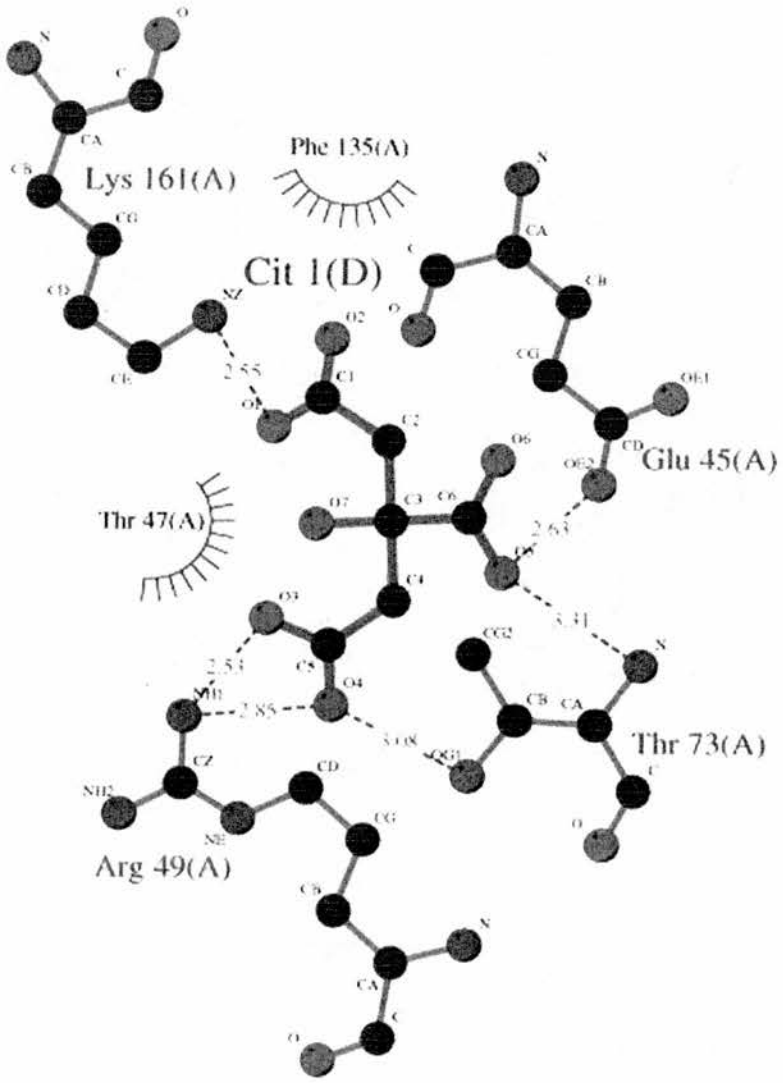


Figure 5.06. The Hydrogen bonding network in monomer A

Conclusions

This is the first instance where directed evolution has successfully relocated the key catalytic residue, specifically onto strand 7 of the β-barrel. This alteration of the location of the Schiff base forming residue results in a concomitant decrease, but not abolition, of activity. This alteration, more importantly, significantly broadens the substrate profile of the enzyme, thus enhancing its utility in organic synthesis. This process of directed evolution may be considered to be in parallel with natural evolution; in natural evolution an enzyme must be able to evolve a different function without losing its original activity. More importantly it provides support for the proposal that one activity can evolve into another through only a few point mutations. These findings, together with those of Fersht, significantly increase the scope for directed evolution to use the α/β barrel as a scaffold for protein engineering experiments.

PDB Accession Code

The KDPG aldolase double mutant has been deposited in the Protein Data Bank. Its accession code is 1FWR.

Chapter 6

Cloning, expression and preliminary characterisation of m62 - a zinc finger protein

6.1 Abstract

m62 contains a novel C-terminal zinc finger like domain. It is novel because the four fingers are comprised of two that use four Cys residues to co-ordinate the zinc ion and two that use three Cys and one His. Normally the fingers are comprised of two Cys and two His residues. It is known that zinc fingers bind DNA and are involved in DNA recognition; they are also required for the correct folding of DNA. It is not fully known how the fingers bind the DNA molecule. Initial studies focused on attempts to clone the m62 DNA into a pET-11c vector. When this proved unsuccessful pure protein was obtained from a pGEX clone constructed by our collaborators; m62 was present as a fusion protein with a GST tag. Dynamic light scattering and preliminary NMR analysis showed the pure, cleaved protein to be incorrectly folded and unstable with respect to aggregation at a protein concentration of 10mg ml^{-1}.

6.2 Introduction

DNA binding proteins are essential for the utilisation of genetic information (Elser *et al.*, 1997). Whilst there are many examples of proteins binding to a specific sequence of double stranded DNA, actual sequence specific interactions with single stranded DNA are thought to occur less frequently. Interaction independent of a specific sequence is another possibility, however only a few sequence independent structural motifs have been described as binding sites (de Murcia & de Murcia, 1994; Weaver, 1995). Transcriptional factors, as a subclass of DNA binding proteins, play an important role during cell growth and differentiation. Several families of transcriptional factors have been described, one of which is the family of zinc finger proteins (Berg & Shi, 1996).

The zinc finger protein was first discovered in studies of the transcriptional control of the 5S ribosomal RNA genes in *Xenopus laevis* in 1985. The gene, known as TFIIIA, contains nine similar domains, each approximately thirty residues in length. Each domain contains two Cys and two His residues; C2H2 (where C is Cys and H is His). These residues are in identical positions in each finger and co-ordinate the zinc in a tetrahedral manner. Each thirty residue repeating unit was found to fold into an elongated zinc binding domain having the shape of a 'finger'; hence the name (Miller *et al.*, 1985). An example of a TFIIIA structure with DNA bound is shown in figure 6.01. It is estimated that there are over 500 zinc finger proteins encoded in the yeast genome; potentially 1% of all mammalian genes encode zinc finger proteins (Hoovers *et al.*, 1992).

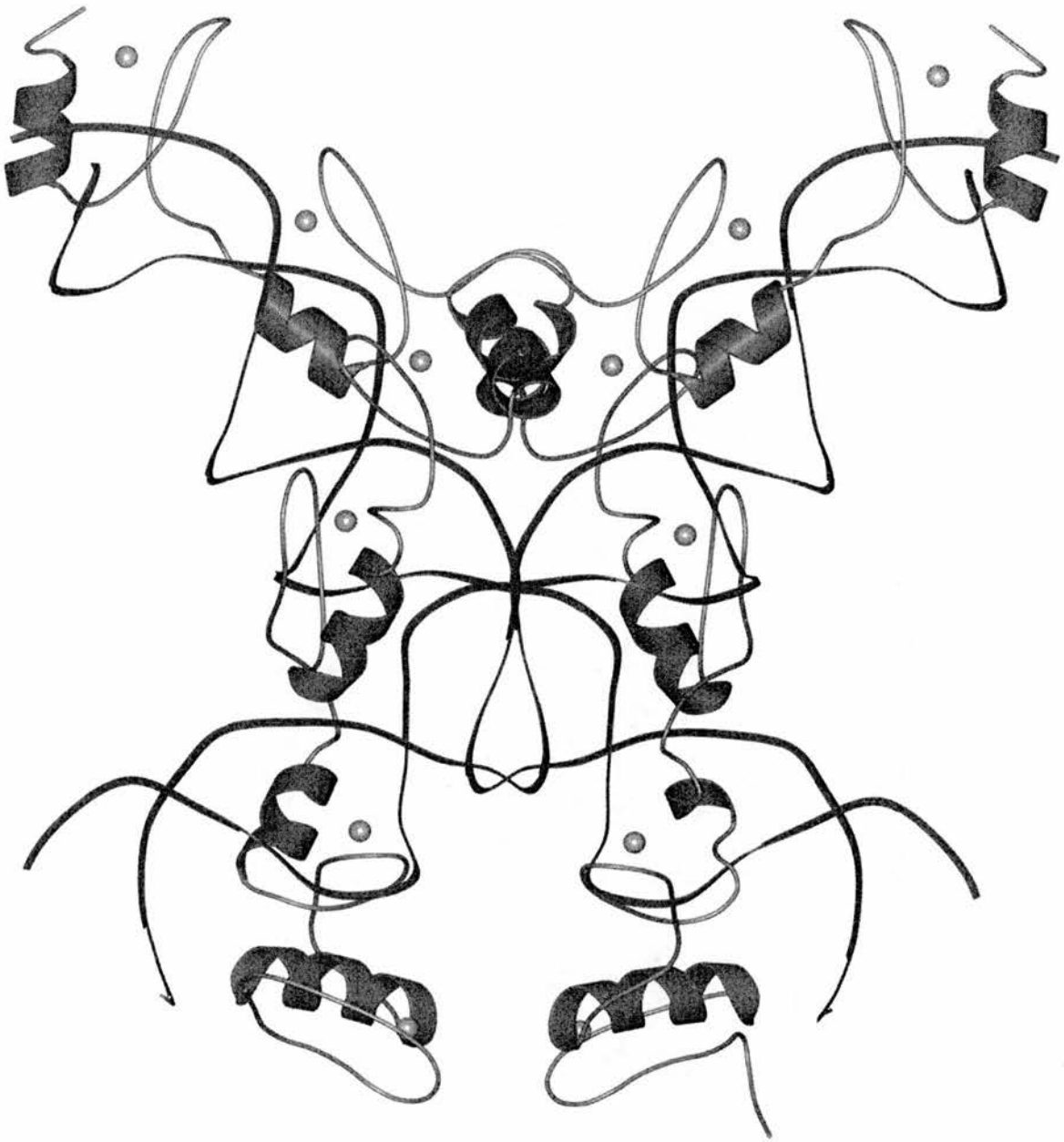


Figure 6.01. TFIIIA with bound DNA. Zinc atoms are shown in yellow, DNA in sky-blue and protein shown in lilac and purple.

As the first sequence specific DNA binding protein to be purified from eukaryotic cells, TFIIIA has become the archetypal zinc finger protein and is an influential model for understanding the mechanisms of site specific DNA and RNA recognition. The number and flexibility of the zinc fingers in TFIIIA, nine in total, has made a complete structural analysis difficult. Work has therefore mainly focused on peptide fragments of TFIIIA containing one to three zinc fingers. This work has shown the N-terminus to fold into a pair of anti-parallel ß strands containing the Zn^{2+} ion co-ordinating Cys, and the C-terminus to fold into a helix containing the conserved His residue (Berg & Shi, 1996). The structures of TFIIIA with DNA bound show binding to be mediated largely through the major groove and via phosphate contacts with the helical region of the zinc finger. The initial solution of the zif 268 structure (Pavletich & Pabo, 1991) indicated that a simple code for zinc finger recognition of DNA may be deducible. More recent structures have suggested that a multiplicity of recognition modes are used by different zinc finger proteins even though the overall structure of the zinc finger domains exhibit considerable similarity.

Although such work on the fragments of TFIIIA has been helpful, it is important to note that the interpretation of these results is strongly dependent on the following postulate; interaction between individual zinc fingers either does not occur or is unimportant. Thus, the structure and function of the protein fragment must be interpreted to accurately reflect those of the same moiety in the context of the full-length protein. Studies on complementary N- and C- terminal truncation mutants, as well as double bond finger disruption mutants, showed the existence of an unfavourable thermodynamic interaction between the nine zinc fingers of TFIIIA when bound to DNA. A possible explanation for this is that the simultaneous binding of DNA to zinc

fingers at opposite ends of the TFIIIA requires energetically unfavourable distortions in the DNA, the protein, or both, relative to the preferred conformation of free DNA and free protein.

Research has shown that the properties of TFIIIA bound to DNA, and particularly the thermodynamic parameters governing the binding equilibrium between TFIIIA and the 5S rRNA gene, cannot be deduced from fragment analysis (Kehres *et al.*, 1997). It has recently been shown that energetic strain is built into the TFIIIA 5S rRNA gene complex (Kehres *et al.*, 1997).

A large subfamily of zinc finger proteins is characterised by the presence of a highly conserved motif (approximately 75 residues) the KRAB-A (*Kruppel* associated box) domain, shown in figure 6.02. This has been identified as a potent transcriptional repressor domain (Pengue *et al.*, 1994). The KRAB-A domain is present in approximately one third of the *Kruppel* type zinc fingers. Once identified, the cloning of proteins which interact with the KRAB-A domain followed. No definitive consensus DNA binding site for a KRAB-zinc finger protein has been identified to date.

$$X_3 - Cys - X_{2\text{-}4} - Cys - X_{12} - His - X_{3\text{-}4} - His - X_4 -$$

Figure 6.02. KRAB-A domain present in all eukaryotes studied to date. This motif is written commonly as C2H2, where C is Cys and H is His.

Studies on a zinc finger protein known as Kid-1 (Elser *et al.*, 1997), a 66kDa protein with thirteen C2H2 fingers at its C-terminus and a KRAB-A domain at its N-terminus,

demonstrated the binding of the zinc finger region to DNA. These studies were the first demonstration of the binding of a KRAB-A zinc finger protein to DNA. It remains unclear how the fingers actually recognise the DNA. Studies on another KRAB-A zinc finger protein, ZNF74 a developmentally expressed gene commonly deleted in DiGeorge syndrome, showed the zinc finger domain to conceal an RNA binding domain *in vitro* (Grondin *et al.*, 1997).

Zinc finger proteins have also become an area of active research as a result of their connection with Human Immunodeficiency Virus 1 (HIV-1). Strategies for the treatment of HIV-1 are currently plagued with the problem of drug resistance. There are two nucleocapsid protein zinc fingers in the virus under investigation, both of which are prime drug targets as they are mutationally intolerant and required for both acute infection and virion assembly. Experiments in which the zinc fingers have been deleted have shown that the new budding virions do not incorporate RNA, which is instead ejected from the infected cell. The nucleocapsid protein and zinc fingers are also thought to play some role during the process of reverse transcription. While the exact role of the protein is not known, it may anchor the RNA molecule into position whilst the reverse transcription enzyme constructs HIV-DNA from the RNA template (Rice *et al.*, 1995).

Zinc finger proteins were identified by sequence analysis of retroviral components, revealing a highly conserved structural motif; the retroviral zinc finger. The chelating residues, three Cys and one His, and the spacing of the zinc finger array are absolutely conserved among all known lentiretroviruses (Gorelick *et al.*, 1990). This is a family of retroviruses that includes HIV-1, HIV-2, cancer causing Human T cell lymphocyte virus

(HTLV), feline immunodeficiency virus and murine leukaemia. It has been found that alterations in just one residue results in inactive protein. If the zinc finger is replaced with a zinc finger found normally in cellular proteins, RNA can still be incorporated into budding virions, yet the resulting virus is non-infectious (Rice *et al.*, 1995).

Research has identified a series of non-toxic, disulfide-substituted benzamides that attack the zinc finger protein. These compounds have been found to inactivate cell free virions, inhibit acute and chronic infections and generally exhibit broad antiretroviral activity (Rice *et al.*, 1995). The compounds were found to be highly synergistic with their antiviral agents and resistance mutants have not been detected so far. The retroviral CCHC (where C is Cys and H is His) zinc finger currently stands as a rare conserved feature against the background of extreme variation characteristic of retroviruses. This indicates that mutational circumvention of reagents that selectively target the retroviral zinc finger may be difficult for the virus to achieve. However, such theories have been advanced previously with respect to other targets only to be proven wrong by experience. Further study should provide a powerful impetus for the utilisation of the zinc finger as an antiretroviral target and thus included in strategies for the development of effective drugs against HIV-1 and other retroviral based diseases.

A further emerging area of research involving zinc finger proteins is the engineering of DNA binding proteins to recognise desired sequences (Greisman & Pabo, 1997). The design of such proteins remains a challenging problem. Zinc finger proteins related to TFIIIA and zif 268 have been found to provide the most versatile framework to date. Modelling calculations, sequence comparisons and phage display have been used to adjust the specificity of an individual finger within a multi-finger protein. Fingers have

also been 'mixed and matched' in the attempt to construct new DNA binding proteins. However, as previously mentioned, it is now known that individual fingers within a protein interact with each other. Previous work assumed each finger could be treated as an independent unit. Whilst these studies provided a useful starting point for further work, as DNA binding models they were found to be ineffectual. As more crystallographic structures of zinc finger-DNA complexes were solved, it became evident that context dependent interactions were important for zinc finger DNA recognition.

One method currently proposed enables the selection strategy to accommodate the context dependant interactions between neighbouring fingers and subsites. This protocol (Greisman & Pabo, 1997) involves using the zif 268 protein, shown in figure 6.03, (Elrod Erickson *et al.*, 1996; Pavletich & Pabo, 1991) as a framework, and gradually extending the new zinc finger protein across the desired nine or ten base pair target site, adding and optimising one finger at a time. Zinc finger proteins have been obtained which not only bind with nano-molar dissociation constants but are also able to discriminate effectively against non-specific DNA. The strategy also ensures the new fingers are always selected in a relevant structural context. As an intact binding site is present at every stage, and because the selections are performed in the context of a growing protein DNA complex, this method readily optimises context dependant interactions between neighbouring fingers and subsites. Thus the protocol naturally selects for fingers which will function effectively and efficiently together (Greisman & Pabo, 1997).

The majority of this work (Greisman & Pabo, 1997) has focused on the design and selection of sequence specific DNA by zinc finger proteins. Whilst the DNA binding properties of these designed proteins have been studied extensively, the structural basis for site specific binding has not been experimentally examined. A co-crystal structure (Kim & Berg, 1996) revealed sequence-specific DNA binding proteins can be designed that operate through remarkably simple hydrogen-bonding networks between contact amino acids and the DNA bases. Natural zinc finger proteins show such a range of modes of interaction with DNA that recognition rules are difficult to develop.

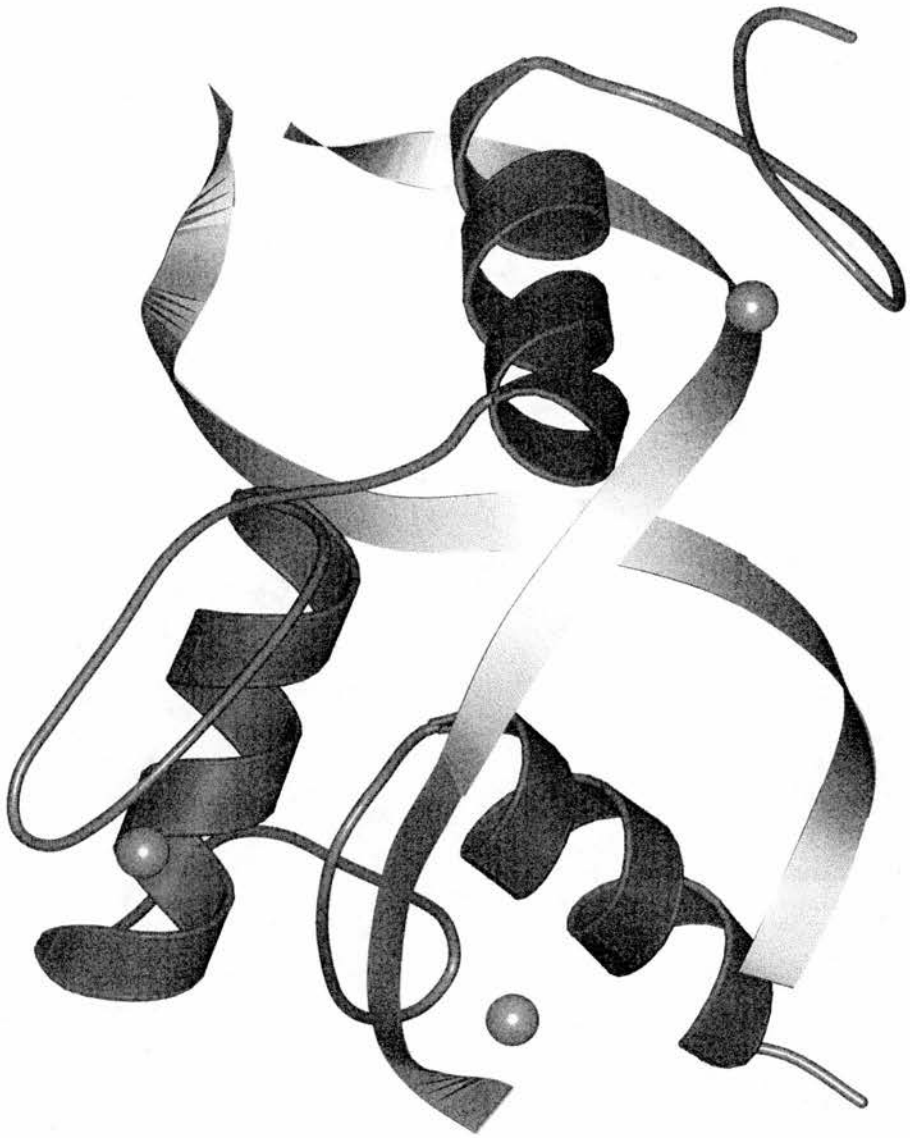


Figure 6.03. Zinc finger protein, Zif268, with bound DNA. Zinc atoms are shown in yellow, DNA double helix in sky-blue, and protein shown in lilac and purple.

neuro-d4

The novel zinc finger protein to be studied during the course of this project, m62, is related to another zinc finger protein, neuro-d4, which was isolated from a screening programme of rat central nervous system cerebral cortex cDNA (Buchman *et al.*, 1992). neuro-d4 was cloned and expressed; its structure, diversity and pattern of expression was then investigated. In contrast to most other previously identified zinc finger genes, neuro-d4 was found to be expressed not only in the nervous system but also in certain neurons.

Sequence alignment studies identified one region of homology between neuro-d4 and known sequences, amino acids 193 to 221 matched exactly with the consensus sequence of the *Kruppel* type zinc finger, shown in figure 6.04. However, unlike other known proteins of this class, which have multiple tandemly repeated *Kruppel* type zinc fingers, there is only one in neuro-d4. Instead, neuro-d4 contains novel Cys/His rich sequences resembling fingers, located in the C-terminal domain.

MKr2 consensus KPYXCXXCGKAFXXXXXLXXHQRIHTGE

Neuro-d4 finger1 KPYVCDI-CGKRYKNRPG-LSYHY-T-HTHEAE

Figure 6.04. Alignment of Kruppel zinc fingers from neuro-d4 (residues 193-221) and mKr2

neuro-d4 is the first example of the combination of a *Kruppel* type zinc finger with another finger-like structure in the same protein; the finger-like structures are separated

by 55 residues. The presence of this new type of Cys/His rich region together with a *Kruppel* type zinc finger, an acidic region and a nuclear localisation signal indicate neuro-d4 may be involved in the regulation of gene activity.

In situ hybridisation studies have shown that neuro-d4 mRNA positive cells are neurons. The level of neuro-d4 was found to be higher during development than in the mature nervous system. The change in this level is thought to reflect either different densities of neuro-d4 mRNA positive neurons in various structures or different levels of neuro-d4 expression in the neurons or, potentially, even both (Buchman *et al.*, 1992).

m62

The zinc finger under examination in this work, m62, contains a C-terminal 'zinc finger like' domain which is completely novel. There are four 'fingers' in total, two of which use four Cys residues to co-ordinate the zinc ion, and two of which use three Cys and one His.

On the loops of the fingers of m62 several polar residues are found, including glutamate, serine and glutamine. It is also thought lysine and arginine may be involved in phosphate binding. Figure 6.05 shows the location of the zinc co-ordinating Cys and His, in red, as well as the residues of interest on the individual fingers. An interesting aside from the crystallographic studies would be site directed mutagenesis studies of these residues to determine the effects they have on the mode of action of the zinc fingers. A diagramatic representation of the actual fingers of m62 is shown in figure 6.06.

MPDGTVIPN

GYCDPCLGGSKKTGCPGDLISCADCGR **Finger 1**

SGHPSCLQPTVNMYAAVRTYRWQCIEC **Finger 2**

KSCSLCGTSENDDQLLPCDDCDR **Finger 3**

GYHMTCLSPPMAEPPEGSWSCHLC **Finger 4**

LRHLKEKASAYITLT

Figure 6.05. The Zinc Fingers of m62

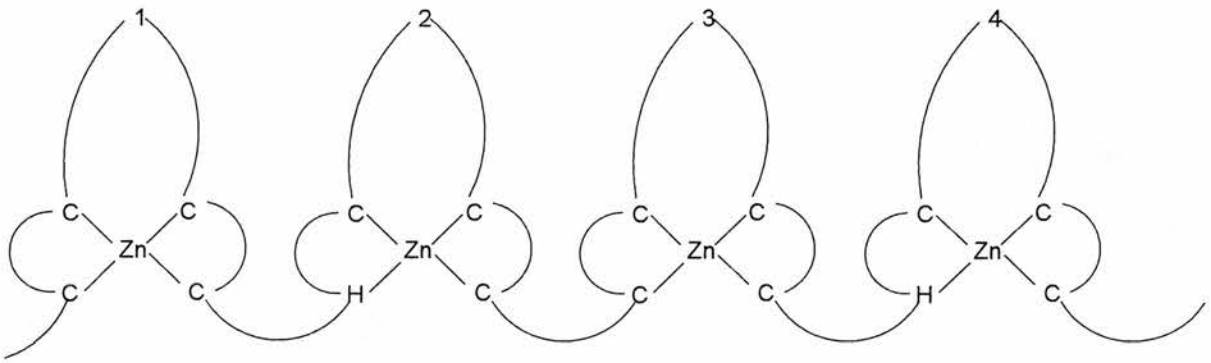


Figure 6.06. The zinc fingers of m62 shown diagrammatically. C is Cys, H is His.

The aim of this work was to clone m62 into the pET-11c expression vector and express the clone in *E. coli* BL21(DE3) cells. Once purified the three-dimensional structure of the protein was to be determined. Further areas of interest are the mutagenesis of residues thought to be essential for DNA binding, micro-calorimetric analysis of m62 binding DNA and crystallographic studies of m62 with bound DNA.

6.3 Experimental

Cloning of m62 into pET-11c

Two oligonucleotide primers were designed according to the target sequence to clone an *NdeI* site into it. The sequence recognition CAT*ATG (where * indicates the position of cutting) was required at the N-terminus for the *NdeI* enzyme and a *BamHI* site, sequence recognition G*GATCC, was required at the C-terminus of the target DNA. The sequences of the two primers are shown in figure 6.07, these were synthesised by Oswell DNA Service. The target DNA was amplified by Polymerase Chain Reaction (PCR); the components of this reaction are shown in table 6.01, the reaction conditions are shown in table 6.02.

N-terminal primer	5' CAG CCA AG**C ATA TG**C CAG ATG GCA CTG
C-terminal primer	5' GCA GAG **GGA TCC** CTA GGT CAG GGT G

Figure 6.07. Oligonucleotides designed for PCR. The recognition sites for the restriction enzymes are shown in blue.

Component	Concentration / Amount
N-terminal Primer	120ρmoles
C-terminal Primer	120ρmoles
Template DNA	50ηg
dNTP's	final concentration 200μM
$MgSO_4$	final concentration 1.25mM
Vent Reaction Buffer	4ul of 10x concentration
Polymerase Enzyme - Vent	1μl
H_2O	to final volume 40μl

Table 6.01. PCR Components

Reaction Step	Temperature (K)	Time (mins)
♦Denaturation	367	1.5
♦Annealing	333	1.5
♦Synthesis	345	2
Termination	345	7

Table 6.02. PCR Reaction Conditions: ♦these steps ran for 35 cycles

The PCR DNA was then washed by the Phenol-Chloroform method. This involved taking the PCR product, adding an equal volume of phenol which was well mixed and centrifuged at 13000g for 4 mins. The top layer was extracted and half the volume of phenol added with half the volume of chloroform, it was mixed again and centrifuged at 13000g for 4 mins. The top layer was extracted and finally an equal volume of chloroform added, mixed and centrifuged at 13000g for 4 mins. The top layer was extracted and the DNA was precipitated by adding two volumes of cold ethanol and

stored at 203K for 1 hour. The sample was centrifuged at 13,000g for 30 mins at 277K. The resultant pellet washed in 70% ethanol, air-dried and resuspended in an appropriate volume of water, typically 50µl.

The insert and pET-11c vector were digested by the enzymes *NdeI* and *BamHI* at 310K for 4 hours. Both the insert and the vector were isolated and purified; they were separated on an agarose gel, the respective bands cut, dissolved in STE buffer and separated by phenol-chloroform extraction.

Once the insert and vector had been digested, an analytical agarose gel was run to determine the concentrations of the insert and vector. This was done to primarily ensure the DNA had not been lost in a previous step and to ensure the correct ratio of insert to vector was used for the ligation reaction. Following this the reaction was prepared as shown in table 6.03 and incubated at 289K overnight. A separate ligation reaction, in which water was used in place of the insert, was also performed to act as a control.

Component	Amount / Concentration
Insert	ratio 10:1 insert to vector
Vector	ratio 10:1 insert to vector
Ligase Reaction Buffer	1µl 10x concentration
Ligase	1µl
Water	to final volume of 10µl

Table 6.03. Ligation Components

After the incubation 5µl of the ligation mix was added to 200µl of *E. coli* XL1 Blue competent cells. They were left on ice for 30 mins, heat shocked at 315K for 42s, plunged back on ice for 2 mins, made up to 1ml with LB and put at 310K for 1 hour. The cells were then centrifuged to concentrate them and resuspended in 100µl of the supernatant, plated out onto Luria Agar plates and incubated at 310K overnight. Analysis 18 hours later showed the presence of colonies. From the plates colonies were selected and grown on a small scale, typically 3-5ml, in LB, 310K, 175 rpm. 16 to 18 hours later the DNA was isolated by Qiagen Mini-Prep Purification Kit (for details see pg200). The DNA was then digested by restriction enzymes, *NdeI* and *BamHI,* which digest the DNA to produce fragments of known sizes, for 2 hours at 310K. The digestions were then run on an analytical agarose gel. The digestions were successful and clearly showed the presence of the insert and vector, shown in figure 6.08.

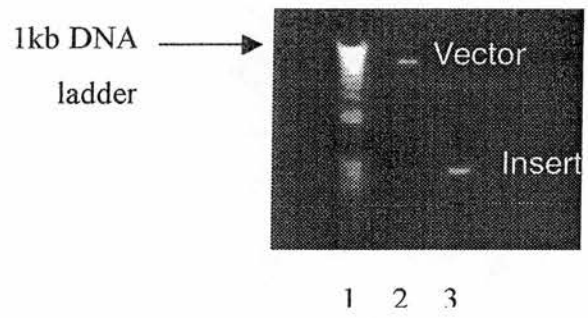


Figure 6.08. Agarose gel of DNA markers, pET vector and the insert. Lane 1 is the DNA 1kb ladder, lane 2 the vector and lane 3 the insert.

The clone was submitted for sequencing DNA Sequencing facility at the University of St Andrews. The components of the sequencing reaction are shown in table 6.04.

Reaction Component	Amount/Concentration
DNA	1.2μg
Sequencing Primers	4ρmole
DMSO	1μl
Water	to final volume 12μl

Table 6.04.Sequencing Reaction Components

Due to its relatively small size m62 only required two sequencing reactions to sequence the entire clone. The details of these primers are in figure 6.09.

P1	N-terminus	GAG CGG ATA ACA ATT TCA CAC AGG
P2	N-terminus	ACC CAT GAC CTT ATT ACC CTC

Figure 6.09. Details of the sequencing primers used for m62 sequencing reactions

The sequencing results showed m62 not to be present in the clone; it was found not to contain the original N- and C-terminal primers. Further trial digestions were examined. The insert was definitely thought to be in the clone, thus it was re-submitted for sequencing. However, the sequencing gave the same result as the first time. Our collaborators had a construct in a pGEX vector, thus it was decided to continue using this clone until the problems with the pET-11c clone could be solved.

pGEX m62 Construct

10ml overnight cultures were grown in LB and used to inoculate 2 litres of LB. The cultures were grown up at 310K, 200 rpm, until an OD_{600} of 0.6 was reached. They were then induced by the addition of IPTG and the cultures incubated under the same

conditions for a further 3 hours. A silver stained SDS-PAGE gel of the over-expression is shown in figure 6.10. The cultures were centrifuged at 8000 rpm, 277K, and resuspended in 10ml of 200mM NaCl, 500mM Tris pH 7.9 and sonicated in 6, 30 s bursts, with 30 s rest time in between, on ice. SDS-PAGE analysis of the cell lysate showed an over-expressed band at approximately 40kDa. This is the combined expected molecular weight of the m62 (12kDa) and the GST tag (26kDa.).

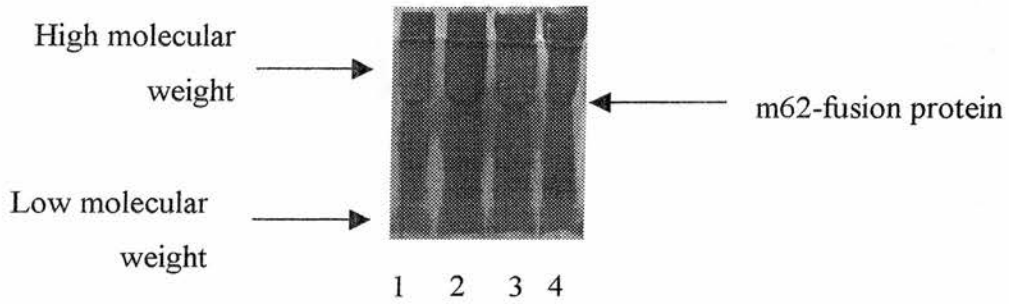


Figure 6.10. SDS PAGE gel of the over-expression of m62-GST tagged protein. Lanes 1-4 show varying concentrations of the sample.

The Glutathione Sepharose 4B was first prepared as a 50% slurry. The beads were then washed using high and low pH solutions, 3 times with 5 mins centrifugation at 500 rpm between each wash. The supernatant was applied to the beads and the mixture incubated at room temperature for 30 mins. The beads were again centrifuged and washed with PBS 3 times; at this stage the solution being removed was analysed by Bradford to ensure the over-expressed protein had bound to the beads and only non specific proteins were being washed off.

The beads were then washed with elution buffer (10mM reduced glutathione, 50mM Tris-HCl) 3 times with centrifugation, 500 rpm for 5 mins. Bradford analysis of the

clutant showed 9.2mg L^{-1} of fusion protein to be obtained. Silver stained SDS-PAGE analysis of the PBS washes and eluted fusion m62 gave non-specific proteins in the former case and a single band at approximately 40kDa in the latter case, shown in figure 6.11.

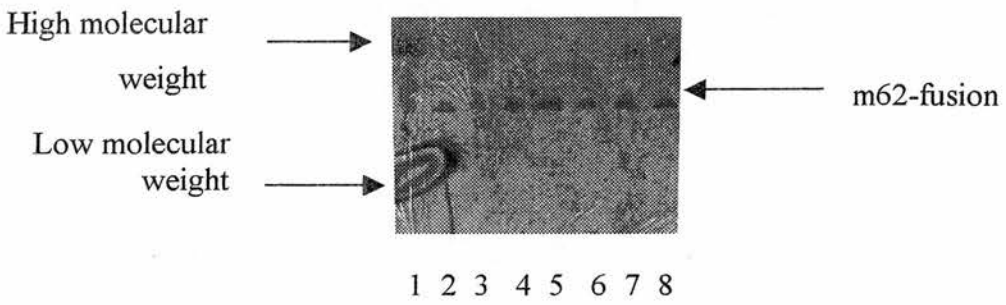


Figure 6.11. SDS PAGE gel of the pure fusion protein eluted from Glutathione sepharose beads. Lane 1 is molecular weight markers, lanes 2-8 pure m62-fusion protein at varying concentrations.

Thrombin Cleavage

Previous experience in our research group had shown that GST-fusion proteins were not amenable to crystallisation. Two methods for cleaving the GST were tested to determine which was more efficient and thus give pure protein. The optimum sites at which thrombin cleaves are shown in figure 6.12.

Site Collection A - **P4-P3-Pro-Arg/Lys •-P1'-P2'** for example -

P4	P3	Pro	R/K	Cut	P1'	P2'
Met	Tyr	Pro	Arg	•	Gly	
Ile	Arg	Pro	Lys	•	Leu	Lys
Leu	***Val***	***Pro***	***Arg***	•	***Gly***	***Ser***

Site Collection B – **P2-Arg/Lys-P1'** for example –

P2	R/K	Cut	P1'
Ala	Arg	•	Gly
Gly	Lys	•	Ala

Figure 6.12. Thrombin cleavage sites

The first method used was cleavage with thrombin on the beads. The crude fusion protein was applied to the beads and incubated at room temperature for 30 mins as before and washed 3 times with PBS. Thrombin was added to the beads at a concentration of 10units mg^{-1} fusion protein. The solution was gently mixed and incubated at room temperature, 295K, with gentle rocking for 2 hours. m62 was eluted in 3 PBS washes and the GST eluted with reduced glutathione. Silver stained SDS-PAGE was used to analyse the purity of the cleaved protein.

The second method was to cleave with thrombin whilst the protein was in solution. In this case the crude protein was purified as before and the fusion protein eluted from the beads. This solution was dialysed thoroughly into PBS to remove all traces of the reduced glutathione. Thrombin was added at a concentration of 10units mg^{-1} fusion protein and the solution was gently mixed and incubated at room temperature, 295K, with gentle rocking for 2 hours. The mixture was applied to the beads to remove GST

and any uncut fusion protein. Silver stained SDS-PAGE was used to analyse the purity of the cleaved protein. Time course experiments to determine the best cleavage conditions were performed. From this, and analysis of protein cut both on and off the beads, it was determined the optimum conditions were cleavage in solution after extensive dialysis into PBS. Thrombin was used at a concentration of 10units mg^{-1} fusion protein at 303K for 12 hours and then incubated at 277K overnight prior to the final purification step.

Structural Studies

Whilst purifying sufficient cut protein to allow crystallisation trials to commence it was decided to undertake initial NMR studies on m62. Its size, 12kDa, suggested it would be amenable to three-dimensional NMR investigations to determine its structure. Preliminary studies were thus made to determine if the m62-GST fusion protein could be over-expressed in Minimal Media. m62 would have to be spin labelled, ^{1}H and ^{13}C, to allow the NMR studies to proceed, the standard method being to use spin labelled compounds in minimal media. The recipe for the Minimal Media is given in table 6.05.

Component	Amount Required / ml L^{-1}
Sterile H_2O	750
5 x M9 Salts *	200
1M $MgSO_4$	2
20% Glucose	20
1M $CaCl_2$	0.1

Table 6.05. Composition of Minimal Media

Component	Amount Required / g L^{-1}
$Na_2HPO_4.7H_2O$	64
KH_2PO_4	15
NaCl	2.5
NH_4Cl	5

Table 6.06. *Composition of M9 salts used in Minimal Media

The media were grown in an identical fashion to LB, however for the cultures to reach an OD_{600} of 0.6 took 18 hours compared with the 3 hours LB media takes. Cells were induced by the addition of IPTG and the cultures left for 3 hours. The cultures were centrifuged, 8000g 15 mins 277K, and resuspended in 10ml of 20mM NaCl, 500mM Tris pH 7.9 and sonicated in 6, 30 s bursts, with 30 s rest time in between, on ice. The cell debris was pelleted and the resulting supernatant analysed by silver stained SDS-PAGE gel. The gel showed over-expression of a protein at 40kDa, the predicted GST-m62 fusion weight, shown in figure 6.13. The fusion protein from the minimal media was purified and cleaved using the previously optimised protocols as described before. Pure, cleaved m62 was obtained at a yield of 2mg L^{-1}.

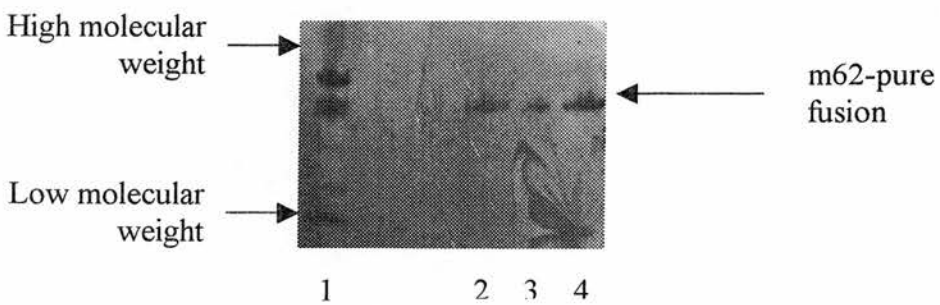


Figure 6.13. SDS PAGE gel of pure fusion protein from minimal media. Lane 1 markers, lanes 2-4 are pure m62-fusion protein at varying concentrations.

Pure protein from LB had meanwhile been concentrated to 10mg ml^{-1} to allow preliminary NMR data to be collected. The initial ^{1}H NMR spectrum, figure 6.14, indicated the protein was not folded. There were also signs that m62 started to precipitate out of solution when at room temperature at this concentration.

Dynamic light scattering was performed on pure m62 at a concentration of 5mg ml^{-1}, the sample was stored on ice until immediately before being used in the experiment. This showed the protein to be composed of ill-defined aggregates.

6.4 Results & Discussion

Attempts to clone m62 into the pET-11c vector were unsuccessful for reasons which remain elusive at this time. The DNA gels of the pseudo-clone clearly showed two bands of the sizes calculated for the m62 insert DNA plus vector. Initially it was thought the problem occurred in the sequencing. However re-sequencing showed the m62 clone was not present in the pET-11c vector. Studies could have continued with the cloning of m62 into the desired pET vector however it was decided to use the pGEX clone isolated by our collaborators.

The pGEX-m62 clone was obtained from Buchman, University of St Andrews. Abundant quantities of fusion protein were obtained from culture in LB. Purification was carried out using Glutathione Sepharose 4B (Pharmacia Biotech). Pure fusion protein, that is GST tagged m62, could be obtained with relative ease. Previous experience had shown fusion proteins would not crystallize. Thus it was decided to cleave the tag using thrombin. One problem often encountered with using thrombin to cleave a tag is the presence of pseudo-sites in the protein attached to the tag. There were no pseudo-thrombin sites in the protein thus full length GST and intact m62 were obtained following cleavage. The optimal cleavage results for this fusion protein were found to be cleaving off the beads after extensive dialysis into PBS. Thrombin, at a concentration of 10units mg^{-1} fusion protein was added to the solution and the resultant incubated at 303K for 12 hours and then incubated at 277K overnight prior to the final application to the beads.

The initial ^{1}H NMR spectrum was not encouraging. The protein, m62, was concentrated to 10mg ml^{-1}; the concentration required for NMR analysis being approximately proportional to the molecular weight of the protein. At this concentration, and whilst at room temperature for the duration of the NMR experiments, approximately 30 mins, the solution in the NMR tube became cloudy; this was attributed to the protein precipitating. The actual ^{1}H NMR spectra obtained showed no evidence of structure. The spectra from 5 to 7 ppm shows no evidence of amide organisation except discrete peaks. From 0 to –1 ppm there is no evidence of any methyl shift having occurred; this shift occurs when aromatic residues stack against methyl groups. Both of these features are characteristic of protein NMR spectra.

Typically light scattering will indicate if the protein is mono-disperse or poly-disperse, and give an indication of its molecular weight in solution. We would have expected to see 24 kDa, 36 kDa or 48 kDa if it were forming monomers, dimers, trimers or tetramers in solution. m62 was found to be poly-dispersed and comprised of ill-defined oligomers with an estimated molecular weight in the range of hundreds of kDa's, and as such be considered non-folded.

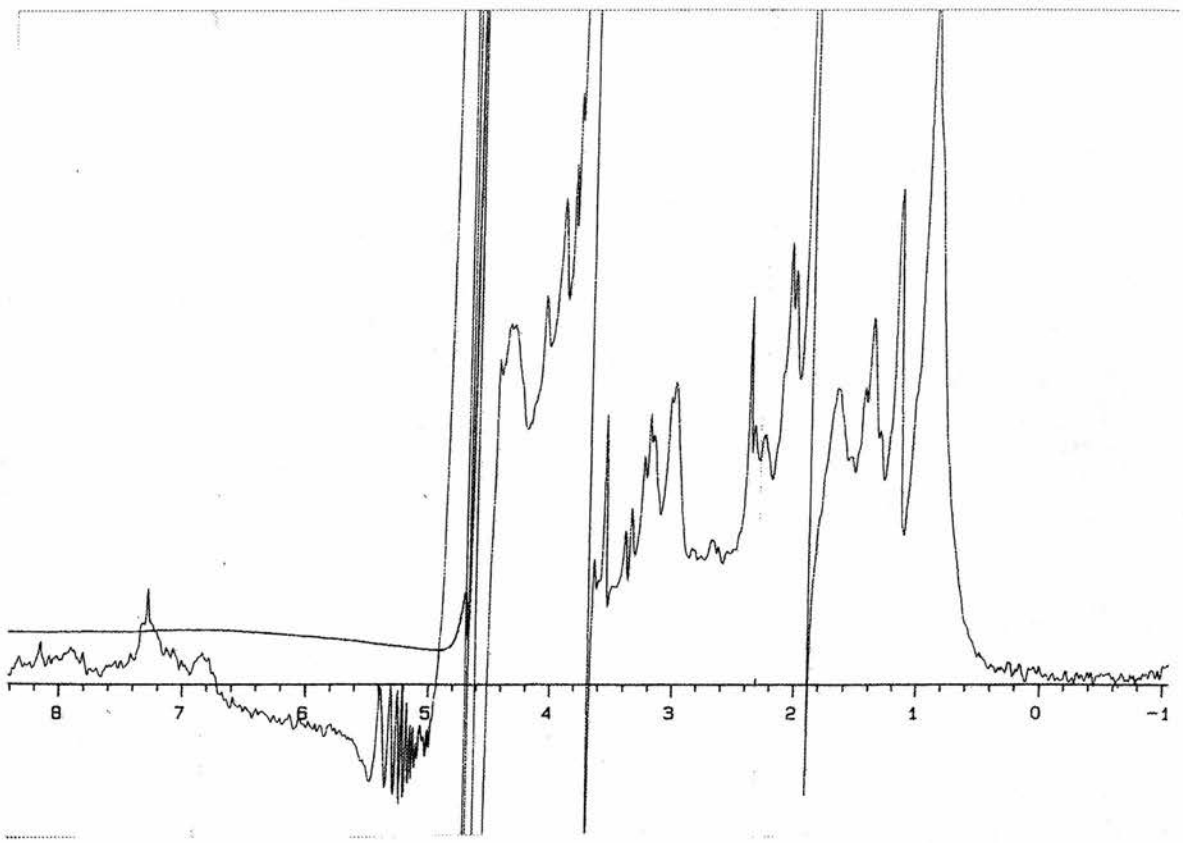


Figure 6.14. ^{1}H NMR spectra of m62

It was at this stage that the project was abandoned. The lack of evidence for folded protein indicated it would not be possible to carry out structural determination studies. It has since been found that to obtain correct folding of zinc finger proteins they should be grown in insect cell lines, such as *baculovirus*, which provide a source of zinc ions now known to be essential for correct protein folding (L. Pearl, personal communication).

Qiagen Mini-Prep Purification Protocol *

1 The pellet from an overnight 3 ml LB culture was resuspended in 250 µl of Buffer P1 and transferred to a microfuge tube.

2 250 µl of Buffer P2 was added and the tube gently inverted 6 times.

3 350 µl of Buffer N3 was added and the tube gently inverted 6 times again.

4 The sample was centrifuged at 13,000 g for 10 mins.

5 The supernatant from the previous step was applied to a QIA prep column and centrifuged at 13,000g for 1 min; the flow-through was discarded.

6 The QIAprep column was washed by adding 750 µl of Buffer PE and again centrifuged at 13,000g for 1 min.

7 The flow-though was discarded and the column re-centrifuged at 13,000g for a further 1 min.

8 The QIAprep column was transferred to a sterile eppendorf tube and the DNA eluted by addition of 50 µl of H_20, left for 1 min and centrifuged at 13,000g for 1 min.

* For further details of the solutions used see the QIAprep Plasmid Handbook.

Bibliography

Allen, S. T., Heintzelman, G. R., and Toone, E. J. (1992) Pyruvate aldolases as reagents for stereospecific aldol condensation. *Journal of Organic Chemistry* **57,** 426-427.

Altamirano, M. M., Blackburn, J. M., Aguayo, C., and Fersht, A. R. (2000) Directed evolution of new catalytic activity using the α/β-barrel scaffold. *Nature* **403,** 617-622.

Banner, D. W., Bloomer, A., Petsko, G., Phillips, D., Pogson, C., Wilson, I., Corran, P., Furth, A., Milman, J., Offord, R., Priddle, J., and Waley, S. G. (1975) Structure of chicken muscle triose-phosphate isomerase determined crystallographically at 2.5Å resolution using amino acid sequence data. *Nature* **255,** 609-614.

Barbas, C. F., Wang, Y.-F., and Wong, C.-H. (1990) Deoxy-ribose-5-phosphate Aldolase as a synthetic catalyst. *Journal of the American Chemical Society* **112,** 2013-2014.

Berg, J., and Shi, Y. (1996) The galvanization of biology: a growing appreciation for the roles of zinc. *Science* **271,** 1081-1085.

Blom, N., and Sygusch, J. (1996) Product binding and the role of the C-terminal region in Class I D-fructose-1,6-bisphosphate aldolase. *Nature Structural Biology* **4** (1), 39-39.

Bradford, M. (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anaytical. Biochemistry* **72,** 248-254.

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

Buchman, V. L., Ninkina, N. N., Bogdanov, Y. D., Bortvin, A. L., Akopian, H. N., Kiselev, S. L., Krylova, O. Y., Anokhin, K. V., and Georgiev, G. P. (1992) Differential splicing creates a diversity of transcripts from a neurospecific developmentally regulated gene encoding a protein with new zinc-finger motifs. *Nucleic Acids Research* **20** (21), 5579-5585.

Budisa, N., Steipe, B., Demange, P., Eckerskorn, C., Kellerman, J., and Huber, R. (1995) High-level biosynthetic substitution of methionine in proteins by analogs 2-aminohexanoic acid, selenomethionine, telluromethionine and ethionine in *Eschericia coli*. *European Journal of Biochemistry* **230** (2), 788-796.

Budisa, N., Karnbrock, W., Steinbacher, S., Humm, A., Prade, L., Neufeind, T., Moroder, L., and Huber, R. (1997) Bioincorporation of Telluromethionine into Proteins: A Promising New Approach for X-ray Structure Analysis of Proteins. *Journal of Molecular Biology* **270** (4), 616-623.

Cadwell, R. C., and Joyce, G. F. (1994) Mutagenic PCR. *PCR Methods Appl.* **3,** S136-140.

CCP4. (1994) The CCP4 suite: programs for protein crystallography. *Acta Crystallographica Section D* **50,** 760-763.

Chou, W.-C., Fotsch, C., and Wong, C.-H. (1995) Synthesis of Nitrocyclitols based on enzymatic aldol reaction and intramolecular nitroaldol reaction. *Journal of Organic Chemistry* **60,** 2916-2917.

Conway, T. (1992) The Entner-Doudoroff pathway: history, physiology and molecular biology. *FEMS Microbial Reviews* **103,** 1-28.

Cooper, S. J., Leonard, G. A., McSweeney, S. M., Thompson, A., Naismith, J. H., Qamar, S., Plater, A. R., Berry, A., and Hunter, W. N. (1996) The crystal structure of a class II fructose-1,6-bisphosphate aldolase shows a novel binuclear metal-binding active site embedded in a familiar fold. *Structure* **4,** 1303-1315.

Cotterill, I. C., Shelton, M. C., Machemer, D. E. W., Henderson, D. P., and Toone, E. J. (1998) Effect of phosphorylation on the reaction rate of unnatural electrophiles with 2-keto-3-deoxy-6-phosphogluconate aldolase. *Journal of the Chemistry Society Perkin Transactions I*, 1335-1341.

Cowtan, K. (1994) *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography*. **31,** 34-38.

Crameri, A., Raillard, S. A., Bermudez, E., and Stemmer, W. P. (1998) DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* **392,** 288-291.

Dalby, A., Dauter, Z., and Littlechild, J. A. (1999) Crystal structure of human muscle aldolase complexed with fructose-1-6-bisphosphate: Mechanistic implications. *Protein Science* **8,** 291-297.

Danson, M. J. (1993) *Central Metabolism of the archaea*. The biochemistry of the archaea (archaebacteria) (Kates, M., Kushner, D. J., and Matheson, A. T., Eds.), Elsvier, Amsterdam

de Murcia, G., and de Murcia, J. (1994) Poly(ADP-ribose) polymerase: a molecular nick-sensor. *Trends in Biochemical Science* **19,** 172-176.

Deacon, A. M., and Ealick, S. E. (1999) Selenium-based MAD phasing: setting the sites on larger structures. *Structure* **7,** R161-R166.

Ding, K., Ishii, A., and Mikami, K. (1999) Super high throughput screening (SHTS) of chiral ligands and activators: asymmetric activation of chiral diol-zinc catalyst by chiral nitrogen activators for the enantioselective addition of diethylzinc to aldehydes. *Angeward Chemistry Int. Ed* **38,** 497-501.

Egan, S. E., Fliege, R., Tong, S., Shibata, A., Wolfe, R. E. J., and Conway, T. (1992) Molecular Characterization of the Entner-Doudoroff Pathway in *Eschericia coli*:

Sequence Analysis and Localization of Promoters for the *edd-eda* Operon. *Journal of Bacteriology* **174** (14), 4638-4646.

Elrod Erickson, M., Rould, M., Nekludova, L., and Pabo, C. (1996) Zif268 protein-DNA complex refined at 1.6Å: a model system for understanding zinc finger-DNA interactions. *Structure* **4,** 1171-1180.

Elser, B., Kriz, W., Bonventre, J. V., Englert, C., and Witzgall, R. (1997) The Krupple-associated box (KRAB)-zinc finger protein Kid-1 and the Wilms' tumour protein WT1, two transcriptional repressor proteins, bind to heteroduplex DNA. *Journal of Biological Chemistry* **272** (44), 27908-27912.

Entner, N., and Doudoroff, M. (1952) Glucose and gluconic acid oxidation of *Pseudomonas saccharophilia*. *Journal of Biological Chemistry* **196,** 853-862.

Erskine, P. T., Senior, N., Awan, S., Lambert, R., Lewis, G., Tickle, I. J., Sarwar, M., Spencer, P., Thomas, P., Warren, M. J., Shoolingin-Jordan, P. M., Wood, S. P., and Cooper, J. B. (1997) X-ray structure of 5-aminolaevulinate dehydratase, a hybrid aldolase. *Nature Structural Biology* **4** (12), 1025-1031.

Fessner, W.-D., and Walter, C. (1996) Enzymatic C-C bond formation in asymmetric synthesis. *Topics in Current Chemistry* **184,** 97-194.

Fessner, W.-D. (1998) Enzyme mediated C-C bond formation. *Current Opinion in Chemical Biology* **2,** 85-97.

Fuhrman, L. K., Wanken, A., Nickerson, K. W., and Conway, T. (1998) Rapid accumulation of intracellular 2-keto-3-deoxy-6-phosphogluconate in an Entner-Doudoroff aldolase mutant results in bacteriostasis. *FEMS Microbial Letters* **159,** 261-266.

Gamblin, S. J., Davies, G., Grimes, J., Jackson, R., Littlechild, J. A., and Watson, H. (1991) Activity and specificity of human aldolases. *Journal of Molecular Biology* **219,** 515-519.

Gijsen, H. J. M., and Wong, C.-H. (1995) Sequential 1 pot aldol reaction catalysed by 2-deoxyribose-5-phosphate aldolase and Fructose-1,6-bisphosphate aldolase. *Journal of the American Chemical Society* **117,** 2947-2948.

Gijsen, H. J. M., Qiao, L., Fitz, W., and Wong, C.-H. (1996) Recent advances in the chemoenzymatic synthesis of carbohydrates and carbohydrate mimetics. *Chemical Reviews* **96,** 48-51.

Gorelick, R., Nigida, S. J., Bess, J. J., Arthur, L., Henderson, L., and Rein, A. (1990) Noninfectious human immunodeficiency virus type 1 mutants deficient in genomic RNA. *Journal of Virology* **64,** 3207-3211.

Greisman, H. A., and Pabo, C. O. (1997) A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science* **275,** 657-661.

Grondin, B., Cote, F., Bazinet, M., Vincent, M., and Aubry, M. (1997) Direct interaction of the KRAB/Cys_2-His_2 zinc finger protein ZNF74 with a hyperphosphorylated form of the RNA polymerase II largest subunit. *Journal of Biological Chemistry* **272** (44), 27877-27885.

Hall, D. R., Leonard, G. A., Reed, C. D., Watt, C. I., Berry, A., and Hunter, W. N. (1999) The crystal structure of *Eschericia coli* Class II Fructose-1,6-bisphosphate Aldolase in complex with phosphoglycolohydroxamate reveals details of mechanism and specificity. *Jounal of Molecular Biology* **287,** 383-394.

Hoovers, J., Mannens, M., John, R., Bliek, J., van Heyningen, V., Porteous, D., Leschot, N., Westerveld, A., and Little, P. (1992) High-resolution localization of 69 potential human zinc finger protein genes: a number are clustered. *Genomics* **12** (2), 254-63.

Izard, T., Lawrence, M., Malby, R., Lilley, G., and Coleman, P. (1994) The three dimensional structure of N-acetylneuraminate lyase from *Eschericia coli. Structure* **2,** 361-369.

Jaeger, K. E., and Reetz, M. T. (2000) Directed evolution of enantioselective enzymes for organic chemistry. *Current Opinion in Chemical Biology* **4,** 68-73.

Jia, J., Huang, W., Schorken, U., Sahm, H., Sprenger, G. A., Lindqvist, Y., and Schnider, G. (1996) Crystal structure of transaldolase B from *Eschericia coli* suggests a circular permutation of the α/β barrel within the class I aldolase family. *Structure* **4,** 715-724.

Jia, J., Schorken, U., Lindqvist, Y., Sprenger, G. A., and Schnider, G. (1997) Crystal structure of the reduced Schiff-base intermediate complex of transaldolase B from *Eschericia coli*: Mechanistic implications for class I aldolases. *Protein Science* **6,** 119-124.

Jones, T., Zou, J., Cowan, S., and M, K. (1991) Improved methods for building protein models in electron-density maps and the location of errors in these models. *Acta Crystallographica Section A* **47,** 110-119.

Kabsch, W. (1976) A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A* **32,** 922-923.

Kehres, D. G., Subramanyan, G. S., Hung, V. S., Roberts, G. W. J., and Setzer, D. R. (1997) Energetically unfavourable interactions among the zinc fingers of transcriptional factor IIIA when bound to the 5S rRNA gene. *Journal of Biological Chemistry* **272** (32), 20152-20161.

Kikuchi, M., Ohnishi, K., and Harayama, S. (2000) An effective family shuffling method using single-stranded DNA. *Gene* **243,** 133-137.

Kim, C. A., and Berg, J. M. (1996) a 2.2Å resolution crystal structure of a designed zinc finger protein bound to DNA. *Nature Structural Biology* **3** (11), 940-945.

Kim, M.-J., Hennen, W. J., Sweers, H. M., and Wong, C.-H. (1988) Enzymes in carbohydrate synthesis: N-Acetylneuraminic acid aldolase catalysed reactions and

preparation of N-Acetyl-2-deoxy-D-neuraminic acid derivatives. *Journal of the American Chemical Society* **110,** 6481-6486.

Kissinger, C., Gehlhaar, D., and Fogel, D. (1999) Rapid automated molecular replacement by evolutionary search. *Acta Crystallographica Section D* **55** (2), 484-91.

Kragl, U., Godde, A., Wandrey, C., Lubin, N., and Auge, C. (1994) New synthetic applications of sialic acid aldolase, a useful catalyst for KDO synthesis, Relation between substrate conformation and enzyme stereoselectivity. *Journal of the Chemical Socity Perkin Trans I* **1,** 119-124.

Lebbink, J. H. G., Kaper, T., Bron, P., van der Oost, J., and de Vos, W. M. (2000) Improving low-temperature catalysis in the hyperthermostable *pyrococcus furiosus* β-Glucosidase CelB by directed evolution. *Biochemistry* **39,** 3656-3665.

Leslie, A. (1992) Joint CCP4 and ESF-EACMB Newsletter on protein crystallography. **26,**44-51.

Leung, D.W., Rice, G.C., Goeddel, D.V., Cachianes, G., Woronicz, J., Chen, E.Y. and Williams, S.R. (1992) Random PCR mutagenesis screening of secreted proteins by direct expression in mammalian cells. *Proc Natl Acad Sci U S A.* **89**, 5467-5471.

Levine, M., Muirhead, H., Stammers, D. K., and Stuart, D. I. (1978) Structure of pyruvate kinase and similarities with other enzymes: possible implications for protein taxonomy and evolution. *Nature* **271,** 626-630.

Lin, C.-H., Sugai, T., Halcomb, R. L., Ichikawa, Y., and Wong, C.-H. (1992) Unusual stereoselectivity in sialic acid aldolase catalysed aldol condensations: Synthesis of both enantiomers of High-Carbon monosaccharides. *Journal of the American Chemical Society* **114,** 10138-10145.

Machajewski, T. D., and Wong, C.-H. (2000) The catalytic asymmetric aldol reaction. *Angeward Chemistry International Edition* **39,** 1352-1374.

Mandecki, W. (1998) The game of chess and searches in protein sequence space. *TIBTECH* **16,** 200-202.

Mavridis, I. E., Hatada, M. H., Tulinsky, A., and Lebioda, L. (1982) Structure of 2-Keto-3-deoxy-6-phosphogluconate Aldolase at 2.8Å Resolution. *Journal of Molecular Biology* **162,** 419-444.

Mavridis, I. M., and Tulinsky, A. (1976) The Folding and Quaternary Structure of Trimeric 2-Keto-3-deoxy-6-phosphogluconic Aldolase at 3.5Å Resolution. *Biochemistry* **15** (20)**,** 4410-4417.

May, O., Nguyen, P. T., and Arnold, F. H. (2000) Inverting enantioselectivity by directed evolution of hydrantoinase for improved production of L-methionine. *Nature Biotechnology* **18,** 317-320.

Merz, A., Yee, M., Szadkowski, H., Pappenberger, G., Crameri, A., Stemmer, W. P. C., Yanofsky, C., and Kirschner, K. (2000) Improving the Catalytic Activity of a Thermophilic Enzyme at Low Temperatures. *Biochemistry* **39,** 880-889.

Miller, J., McLachlan, A., and Klug, A. (1985) Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. *EMBO Journal* **4,** 1609-1614.

Milne, A. A. (1928) *The House at Pooh Corner.*

Miyazaki, K., and Arnold, F. H. (1999) Exploring nonnatural evolutionary pathways by saturation mutagenesis: rapid improvement of protein function. *Journal of Molecular Evolution* **49,** 716-720.

Navaza, J., and Saludjian, P. (1997) AMoRe: An automated molecular replacement program package. *Methods in Enzymology* **276,** 581-594.

Nixon, A. E., and Firestine, S. M. (2000) Rational and "Irrational" Design of Proteins and Their Use in Biotechnology. *IUBMB Life* **49,** 181-187.

Otwinowski, Z. (1996) in *Methods in Enzymology* (Carter, C., and Sweet, R., eds) Vol. 276, pp. 307-326, Academic Press

Pavletich, N., and Pabo, C. (1991) Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1Å. *Science* **252,** 809-817.

Peekhaus, N., and Conway, T. (1998) What's for Dinner?: Entner-Doudoroff Metabolism in *Eschericia coli*. *Journal of Bacteriology* **180** (14), 3495-3502.

Pengue, G., Calabrò, V., Bartoli, P., Pagliuca, A., and Lania, L. (1994) Repression of transcriptional activity at a distance by the evolutionarily conserved KRAB domain present in a subfamily of zinc finger proteins. *Nucleic Acids Research* **22,** 2908-2914.

Plater, A. R., Zgiby, S. M., Thomson, G. J., Qamar, S., Wharton, C. W., and Berry, A. (1999) Conserved residues in the mechanism of the E. coli Class II FBP-aldolase. *Jounal of Molecular Biology* **285,** 843-855.

Ramakrishnan, C., and Ramachandran, G. (1965) Stereochemical criteria for polypeptide and protein chair conformation. *Biophysical Journal* **5,** 909-933.

Reetz, M. T., Becker, M. H., Kuhling, K. M., and Holzwarth, A. (1998) Time resolved IR-thermographic detection and screening of enantioselectivity in catalytic reactions. *Angeward Chemistry International Edition* **37** (2647-2650)

Reetz, M. T., Becker, M. H., Klein, H.-W., and Stockigt, D. (1999) A method for high-throughput screening of enantioselective catalysts. *Angeward Chemistry International Edition* **38,** 1758-1761.

Reetz, M. T., and Jaeger, K. E. (1999) Superior Biocatalysts by Directed Evolution. *Topics in Current Chemistry* **200,** 31-57.

Reetz, M. T., and Jaeger, K. E. (2000) Enantioselective Enzymes for Organic Synthesis Created by Directed Evolution. *Chemistry a European Journal* **6** (3), 407-412.

Rice, W. G., Supko, J. G., Malspesis, L., Buckheit, R. W. J., Clanton, D., Bu, M., Graham, L., Schaeffer, C. A., Turpin, J. A., Domagala, J., Gogliotti, R., Bader, J. P., Halliday, S. M., Coren, L., Sowder, R. C. I., Arthur, L. O., and Henderson, L. E. (1995) Inhibitors of HIV nucleocapsid protein zinc fingers as candidates for the treatment of AIDS. *Science* **270,** 1194-1197.

Richardson, J. S. (1979) The singly-wound parallel β–barrel: a proposed structure for 2-keto-3-deoxy-6-phosphogluconate aldolase. *Biochemical and Biophysical Research Communications* **90** (1), 285-290.

Romano, A. H., and Conway, T. (1996) Evolution of carbohydrate metabolic pathways. *14th Forum in Microbiology*, 448-455.

Shao, Z., Zhao, H., Giver, L., and Arnold, F. H. (1998) Random-priming in vitro recombination: an effective tool for directed evolution. *Nucleic Acids Research* **26,** 681-683.

Shelton, M. C., Cotterill, I. C., Novak, S. T. A., Poonawala, R. M., Sudarshan, S., and Toone, E. J. (1996) 2-Keto-3-deoxy-6-phosphogluconate aldolases as catalysts for stereocontrolled carbon-carbon bond formation. *Journal of the American Chemical Society* **118** (9), 2117-2125.

Shibata, K., Shingu, K., Vassilev, V. P., Nishide, K., Fujita, T., Node, M., Kajimoto, T., and Wong, C.-H. (1996) Kinetic and thermodynamic control of L-threonine aldolase catalysed reaction and its application to the synthesis of mycestericin D. *Tetrahedron Letters* **37,** 2791-2794.

Smith, J. L., and Thompson, A. (1998) Reactivity of selenomethionine - dents in the magic bullet? *Structure* **6,** 815-819.

Spiller, B., Gershenson, A., Arnold, F. H., and Stevens, R. C. (1999) A structural view of evolutionary divergence. *Biochemistry* **96** (22), 12305-12310.

Steipe, B. (1999) Evolutionary Approaches to Protein Engineering. *Current Topics in Microbiology & Immunology*, 55-86.

Stemmer, W. P. (1994) DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proceedings of the National Academy of Sciences USA* **91,** 10747-10751.

Sugai, T., Shen, G.-J., Ichikawa, Y., and Wong, C.-H. (1993) Synthesis of 3-Deoxy-D-Manno-2-octulosonic acid (KDO) and its analogues based on KDO aldolase catalysed reactions. *Journal of the American Chemical Society* **115,** 413-421.

Suzuki, N., and Willis, A. W. (1980) Complete primary structure of 2-Keto-3-deoxy-6-phosphogluconate Aldolase. *Journal of Biological Chemistry* **255** (8), 3427-3435.

Taha, T. S. M., and Deits, T. L. (1994) Purification and characterisation of 2-keto-3-deoxy-6-phosphogluconate aldolase from *Azobacter vinelandii*: Evidence that the enzyme is bifunctional towards 2-keto-4-hydrozy glutarate cleavage. *Biochemical and Biophysical Research Communications* **200** (1), 459-466.

Terwilliger, T. C., and Berendzen, J. (1999) Automated MAD and MIR structure solution. *Acta Crystallographica Section D* **D55,** 849-861.

Uchiyama, V., Vassilev, T., Kajimoto, T., Wong, W., Huang, H., Lin, C. C., and Wong, C.-H. (1995) Design and synthesis of sialyl LewisX mimetics. *Journal of the American Chemical Society* **117,** 5395-5396.

Vandlen, R. L., Ersfeld, D. L., Tulinsky, A., and Wood, W. A. (1973) Confirmation of a trimeric subunit arrangement for 2-Keto-3-deoxy-6-phosphogluconic aldolase using X-ray crystallographic methods. *Journal of Biological Chemistry* **248** (6), 2251-2253.

Vassilev, T., Uchiyama, V., Kajimoto, T., and Wong, C.-H. (1995) An efficient chemoenzymatic synthesis of alpha-amino-beta-hydroxy-gamma-butyrolacetone. *Tetrahedron Letters* **36,** 5063-5064.

Verlinde, C. L. M. J., and Quigley, P. M. (1999) Structure-based re-evaluation of the mechanism of Class I Fructose-1,6-bisphosphate Aldolase. *Journal of Molecular Modelling* **5,** 37-45.

Weaver, D. (1995) What to do at an end: DNA double-strand-break repair. *Trends Genetics* **11,** 388-392.

Wong, C.-H., Moris-Varas, F., Hung, S. C., Marron, T. G., Lin, C.-C., Gong, K., and Weitz-Schmidt, G. (1997) Small molecules as structural and functional mimics of sialyl Lewis X tetrasaccharide in selection inhibition: A remarkable enhancement of inhibition by additional negative charge and / or hydrophobic group. *Journal of the American Chemical Society* **119,** 8152-8158.

Zhao, H., Giver, L., Shao, Z., Affholter, J. A., and Arnold, F. H. (1998) Molecular evolution by staggered extension process (StEP) in vitro recombination. *Nature Biotechnology* **16,** 258-261.