

University of Bath



PHD

The crystal structure of glucose dehydrogenase from a thermophilic archaeon

Rossjohn, Jamie

Award date:
1994

Awarding institution:
University of Bath

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 22. May. 2019

THE CRYSTAL STRUCTURE OF GLUCOSE DEHYDROGENASE FROM A THERMOPHILIC ARCHAEON

Submitted by Jamie Rossjohn

for the degree of Ph.D. of the

University of Bath, 1994

COPYRIGHT

Attention is drawn to the fact that copyright of this thesis rests with its author. This copy of the thesis has been supplied on the condition that anyone who consults it is understood to recognise that its copyright rests with the author and that no quotation from the thesis and no information derived from it may be published without the prior written consent of the author.

This thesis may be made available for consultation within the University Library and may be photocopied or lent to other libraries for the purposes of consultation.

A handwritten signature in black ink, appearing to read 'Jamie Rossjohn', with a large, stylized flourish at the end.

Jamie Rossjohn

UMI Number: U550052

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI U550052

Published by ProQuest LLC 2014. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

UNIVERSITY OF BATH LIBRARY		
26	30 JAN 1995	
Ph D		

5088422

Abstract

The X-ray crystal structure of glucose dehydrogenase (GDH) from *Thermoplasma acidophilum*, a homo-tetrameric enzyme that preferentially utilises NADP, was solved by the single isomorphous replacement technique in conjunction with non-crystallographic symmetry averaging. The averaging technique was novel in that the initial crude spherical mask was gradually transformed into a mask that enveloped the complete GDH tetramer. The structure has been refined to a crystallographic R-factor of 19.3% for all reflections between 8 and 2.9Å.

The monomer is comprised of two domains: a centrally located nucleotide-binding domain and a catalytic domain formed from both N- and C-terminal regions of the polypeptide. The nucleotide binding domain of GDH is structurally homologous to that of previously reported dehydrogenases and thus forms the basis of the discussion on the observed coenzyme specificity of GDH.

Unexpectedly, the catalytic domain of GDH shares extensive structural homology to that of the dimeric horse liver alcohol dehydrogenase (LADH), including the possession of a catalytic and structural zinc. Subsequently, this structural homology was strengthened by biochemical studies.

The quaternary structure of GDH can be described as a dimer of LADH dimers, in which the carboxyl ends of the Rossmann fold are the main sites of subunit interaction. The structural lobe of GDH was the other site of interaction.

A structurally based multiple sequence alignment of the long chain, zinc-requiring alcohol/polyol dehydrogenase family was built, from which the mode of substrate binding in GDH was deduced, and a phylogenetic tree constructed.

A comparative analysis using LADH and a number of nucleotide binding domains, suggests that the number of aromatic interactions, and the site of subunit interactions are putative thermophilic features of GDH. In addition the concept of an Archaeal minimal functional unit is proposed.

Acknowledgements

I would like to start with my main Supervisor and friend, Garry, or as he preferred to be known, Master. I felt that I was fortunate to have such an understanding Supervisor, who guided me well and tolerated my sense of humour. I shall never forget how he heroically rescued me from the clutches of the snoring Czech whilst in China. The remainder of the China experience cannot be printed for reasons concerning my safety (and it would upset Margaret). I would also like to thank my two other supervisors, Dave, for his steady and solid support, and Mike for the enthusiastic approach. Susan, my immediate work neighbour, and fellow come-rain-or-shine cyclist, aided me immensely in solving the structure, and I particularly admired her well balanced and reasoned approach to solving annoying computer problems. In the same vein, I would like to thank Peter for graphical and PC enlightenment. The two other members of the lab., Rupert, co-Archaeal newsletter journalist, and Andrew, refinement and garlic boy, also made for an enjoyable environment. In addition, E.Garman, A.Carver, P. Calleja, and R. Maytum, assisted me with various aspects of my work. I would also like to mention Llyr, who provided me with a computing platform at home, and Lisa, my wife, for supporting me during my time of writing-up and the J.B.A.

CONTENTS	Page
Abstract	i
Acknowledgements	ii
Chapter 1 INTRODUCTION	1-22
Overview	1
The Archaea	1
<i>Thermoplasma acidophilum</i>	6
The enzyme	7
The nucleotide-binding domain	9
Coenzyme specificity	11
Mode and action of nucleotide binding	14
Alcohol dehydrogenase	15
Horse liver alcohol dehydrogenase (and the models)	17
Aims of project	22
Chapter 2 PURIFICATION AND CRYSTALLISATION	23-30
Introduction	23
Bigger is better	25
Results	26
Crystallisation with cofactor	29
Chapter 3 NATIVE DATA COLLECTION AND ANALYSIS	31-46
Introduction	31
In-house detector	32
Data collection and analysis (Area detector)	34
X-ray synchrotron	36
Data collection and analysis (Imaging plate)	37
The unit cell	38

In-house data	39
Synchrotron data	40
Final data file	42
Non-crystallographic symmetry	43
Chapter 4 ISOMORPHOUS REPLACEMENT	47-65
Introduction	47
Heavy atom derivatives	48
The Screen	52
Deleterious derivatives	53
Potential derivatives	53
The successful derivative (the platinum derivative)	58
Derivative analysis	60
Anomalous patterson map	62
Phase calculation	63
Chapter 5 RAVING	66-90
Real space introduction	66
Approximate boundary delineation	66
Non-crystallographic symmetry operators	67
Symmetry averaging introduction	71
The RAVE scheme	73
The mask	75
Averaging within the sphere	76
The Alcohol dehydrogenase mask	82
Phase extension	85
The GDH mask (mark 1)	87
The GDH mask (mark 2)	87
Chapter 6 MODEL BUILDING AND REFINEMENT	91-116

Introduction	91
Topology determination	92
Assignment of C α positions	95
Fitting the primary sequence	95
Refinement	97
NCS restraints	98
Temperature factor refinement	99
Refinement round 1	99
Rounds 2 + 3	100
Round 4	101
Round 5	102
Rounds 6 + 7 + 8	103
Assessment of model quality	104
Heavy atom site	109
Crystal packing	109
Chapter 7 THE GDH STRUCTURE	117-155
Structural characterisation of the monomer	117
The nucleotide binding domain	117
Putative minimal functional unit	127
The catalytic domain	127
The Zinc sites	132
The catalytic zinc	133
The structural zinc	134
The Quaternary structure	136
Multiple sequence alignments	141
Structural framework	141
Structure-function relationships	150
Phylogenetic concerns	154

Chapter 8	COENZYME AND SUBSTRATE INTERACTIONS	156-178
	Introduction	156
	Nucleotide interactions	156
	The GxGxxG/A loop	162
	The hydrogen-bonding network	164
	Substrate binding	165
	The initial docking	166
	The minimised complex	169
	Substrate specificity	172
	GRASP	174
	Continuing the modelling theme - the closed form	174
Chapter 9	THERMOSTABILITY	179-195
	Introduction	179
	Compositional differences	182
	Stabilising helix capping residues	183
	Aromatic/aromatic interactions	184
	The dimeric subunit/subunit interface	190
	Cavities	193
	Putative Archaeal features - the minimal functional unit	194
Chapter 10	BIOCHEMICAL STUDIES	196-209
	Introduction	196
	Determination of zinc	196
	Determination of mechanism	196
	The steady-state kinetic constants	201
	The conformational change	202
	GDH has more similarity to YADH than LADH	206
	Thermostability studies	206

The chelating agent	207
Stability assays	207
Future work	210-213
Appendix 1	214-224
Protein methods	214
Crystallisations	215
Biochemical methods	216
Zinc determination	218
Thermostability studies	219
Phylogenetic tree construction	219
Molecular modelling	219
Substrate/NAD structures	220
Programs/software used	221
Materials	224
Appendix 2	225-233
The unit cell	225
Symmetry within the GDH unit cell	225
Braggs Law (real space)	226
Braggs Law (reciprocal space)	226
The Fourier series/transform	228
Isomorphous replacement	230
Refinement	232
Abbreviations	234
References	235-241

Chapter 1

INTRODUCTION

Overview

The enzyme under investigation is glucose dehydrogenase from the thermophilic Archaeon *Thermoplasma acidophilum*. To begin with, the concept of the Archaea is described, followed by the role played by glucose dehydrogenase. A relatively thorough structural account of the Rossmann fold, factors governing coenzyme specificity and horse liver alcohol dehydrogenase is given with the *a priori* knowledge that glucose dehydrogenase is structurally homologous to the above mentioned enzyme.

The Archaea

As a result of molecular biological studies, the dichotomous nature of life, which was previously based on phenotypic characteristics only, has been rejected; it's overhauling was due to the discovery of a third distinct domain of organisms - the Archaea (archaeobacteria) (Woese, 1990). The discovery of the Archaea and the resultant phylogenetic trees have arisen from studying the relationship between highly conserved and ubiquitous components (Olsen & Woese, 1993), such as RNA polymerase (Iwabe *et al.*, 1991, elongation factor G (Cammarano *et al.*, 1992), proton translocating ATPases (Gogarten *et al.*, 1992) and notably the small subunit ribosomal RNA (Olsen

& Woese, 1993) (3,4), across a wide range of organisms. With the emergence of the Archaea, came a change in the taxonomy. Based on the 16S rRNA data, 3 domains of life have been proposed (fig. 1a): Bacteria (formerly eubacteria), Eucarya (formerly eukaryotes) and the Archaea (formerly archaebacteria) (Woese, 1990). This tree reveals that the Archaea are an evolutionary distinct group of organisms, in which the Archaea and Eucarya share a common lineage before their separation. Furthermore the Archaea are believed to represent the earliest form of life. This is mirrored by their environmental habitats, of which there are three phenotypic classes within the Archaeal domain:

- The halophiles: are adapted to live in extreme saline environments (Larsen, 1973), with their intracellular salt concentrations equating to 5M potassium ions.
- The methanogens are characterised by their ability to reduce carbon dioxide, or simple C1 compounds such as methanol or formate to methane, and by the unusual coenzymes they possess (De Rosa, 1977)
- The sulphur dependent extreme thermophiles are adapted to live at high temperatures (55 → 110°C) and utilise a sulphur acceptor (Stetter & Zillig, 1985).

The taxonomic level below the domain is termed the kingdom. The Archaeal domain is subdivided into two major kingdoms : the Crenarchaeota and the Euryarchaeota (fig. 1b). The former comprise the sulphur dependent extreme thermophiles only, whereas the latter possess all three phenotypes.

However, there isn't complete harmony within the scientific field in regards to the phylogenetic tree. Contradictory results have arisen mainly from Lake (1991, Rivera and Lake, 1992) because different regions of the 16s rRNA sequence have been used to

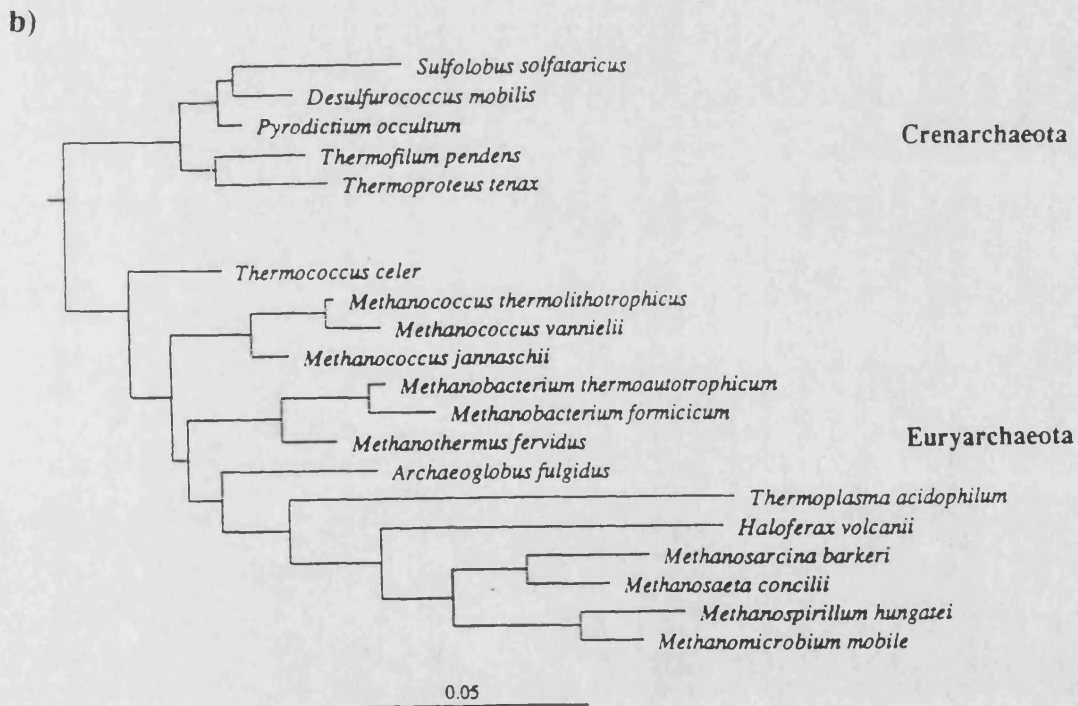
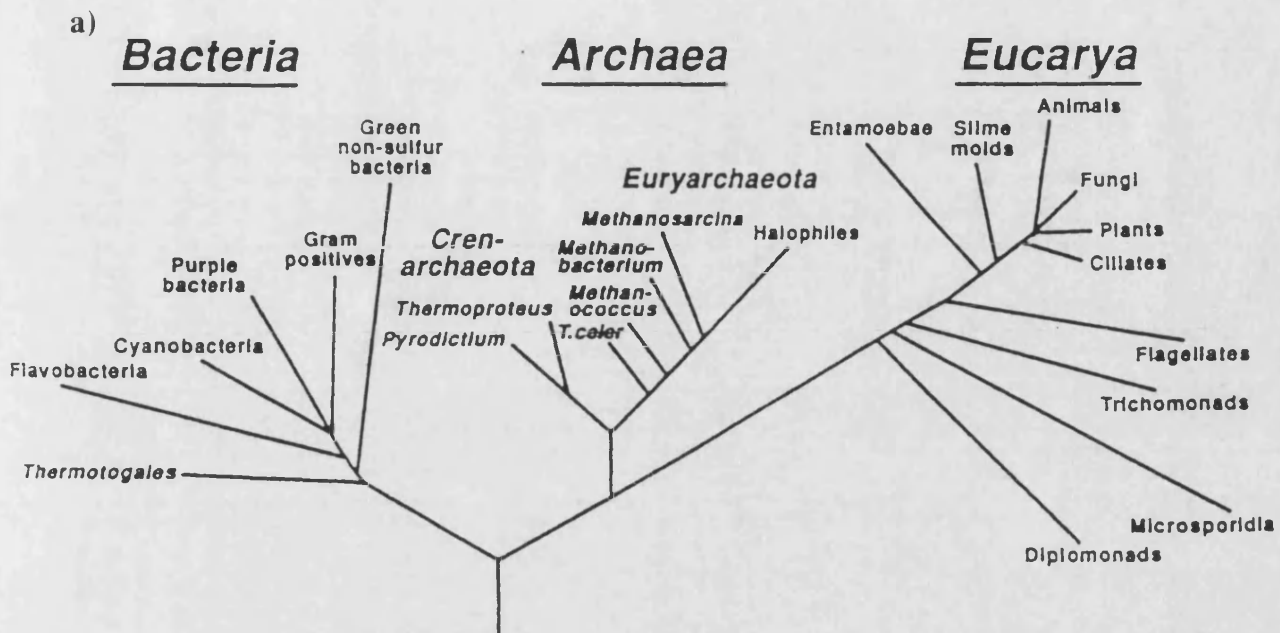


Fig1a) Woese's Archaeal tree. **B)** Divisions within the Archaeal domain - the Crenarchaeota and Euryarchaeota kingdoms.

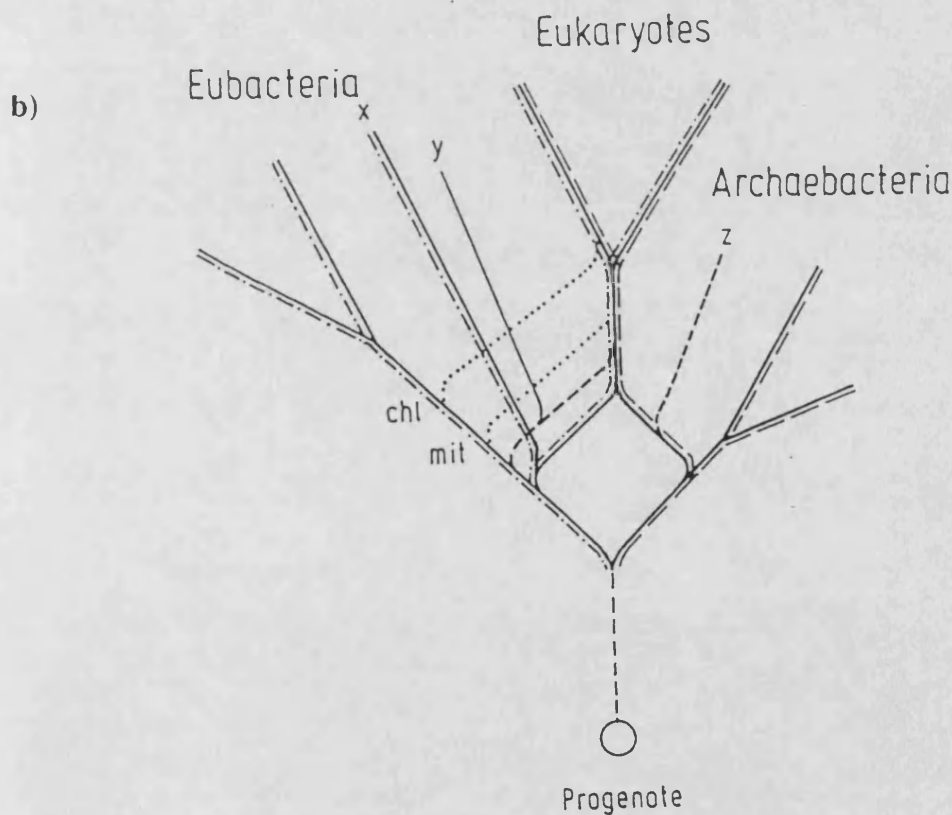
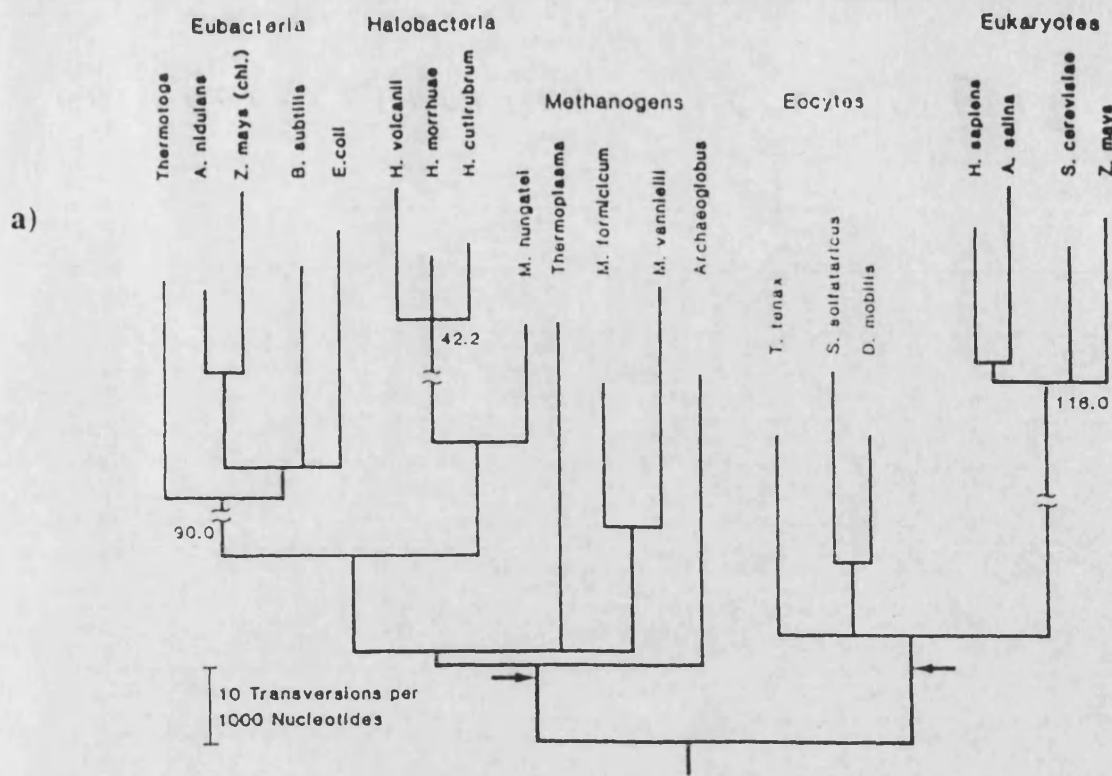


Fig. 1.2 Alternative views on the phylogenetic tree: a) Lake's Eocyte tree. b) Zillig's fusion hypothesis (..... acquisition of mitochondria and chloroplasts; xyz molecules branching from eukaryotic lineages; - - - - -, · - · - · molecules inherited by archaeobacterial and eubacterial ancestors respectively).

construct phylogenetic trees, plus there is more than one way to construct a phylogenetic tree from a given set of data. Distance matrix, maximum likelihood and parsimony are three such methods, with the contradictory results arising from mutational rate biases (unequal rate effects). Lake's phylogenetic tree, the eocyte tree, disputes the existence of a distinct Archaeal domain, but groups the thermophiles to the eukaryotes whilst the halophiles and methanogens are shown to be related to the eubacteria (fig. 1.2a). However, Woese has attributed Lake's tree as a function of the alignment used (as opposed to the method used for constructing the tree), as it is biased towards Lake's tree and includes structurally divergent regions within the alignment. In addition, there is little evidence to support this tree (Olsen & Woese, 1993).

Alternatively, a recent development is the Zillig fusion hypothesis (Zillig *et al.*, 1989) (fig. 1.2a) which, based on DNA-dependent RNA polymerase studies, suggests that the eukaryotes have evolved via a fusion event between the eubacteria and archaeobacteria.

All this contradictory phylogenetic and taxonomic work must not cloud the issue: the Archaea are a unique set of organisms, and the understanding of which will no doubt give a novel insight into how the cellular components can withstand extreme environmental conditions, from which novel biotechnological applications may be found (Hough & Danson, 1989).

Thermophilic and halophilic proteins, and novel Archaeal membrane lipids are three such components that have attracted much interest. Factors governing thermostability are discussed in more detail in Chapter 9. Although no halophilic three-dimensional structures have been reported, models of halophilic enzymes based on their mesophilic

counterparts have led to the suggestion that the former possess a multitude of charged residues on the surface of the protein (Bohm & Jaenicke, 1994). These charged residues would serve to attract water to the surface of the enzyme and counteract the water-withdrawing effect of the high salt concentrations found within the halophilic organism.

Archaeal cell walls consist of glycosylated proteins rather than the peptidoglycan structure in the bacteria, while the Archaeal membrane lipids are unique in consisting entirely of ether lipids (as opposed to ester lipids), which introduces rigidity into the membrane (Kates, 1993).

Despite the fact that cellular and biochemical characteristics should not be used to construct phylogenetic trees, novel features such as novel lipids/cell wall structures (Kandler & Honig, 1993), Archaeal metabolic pathways (Danson, 1993), rRNA splicing enzymes (Garret *et al.*, 1993), that are exclusive to the Archaea serve to cement the existence of the Archaea as a distinct domain.

Thermoplasma acidophilum

Tp. acidophilum was first isolated from a coal refuse pile at the Friar Tuck mine in south-western Indiana (Darland *et al.*, 1970). The organism cultured from this source grows optimally at pH range of 1-2.5 (however, maintains an intracellular pH of 7.0 via a membranous electron transport chain (Searcy, 1986)) and a temperature of 57°C. Initially from cellular morphological studies and biochemical studies *Tp. acidophilum* was classified as a thermoacidophilic prokaryote, however molecular biological studies on ribosomal RNA (Woese & Olsen, 1986) revealed that the *Thermoplasma* belonged to the Archaeal domain (later confirmed by biophysical and biochemical

analyses of ribosomes (Cammarano *et al.*, 1986). These rRNA studies showed that *Tp. acidophilum* belongs to the methanogen - halophile branch, even though it is phenotypically a thermophilic Archaeon.

The true phylogenetic relationship of *Tp. acidophilum* to other Archaea is uncertain, however its position should be clarified by further sequence comparisons.

The enzyme

Differences between the Eucarya, Bacteria and the Archaea are found in the catabolism of glucose (fig. 1.3)(Danson, 1993). The Eucarya and many anaerobic Bacteria proceed via the Embden-Meyerhoff pathway, whereas many strictly aerobic Bacteria lack the enzyme 6-phosphofructokinase, and catabolise glucose via the Entner-Doudoroff pathway. To date, 6-phosphofructokinase has not been detected in any Archaeal organisms. Glucose catabolism in the Archaea is believed to proceed via modifications of the Entner-Doudoroff pathway (the non-phosphorylated Entner-Doudoroff pathway), with the halophiles exhibiting slightly different variations when compared to the pathways of *Thermoplasma* and *Sulfolobus* (Danson, 1993).

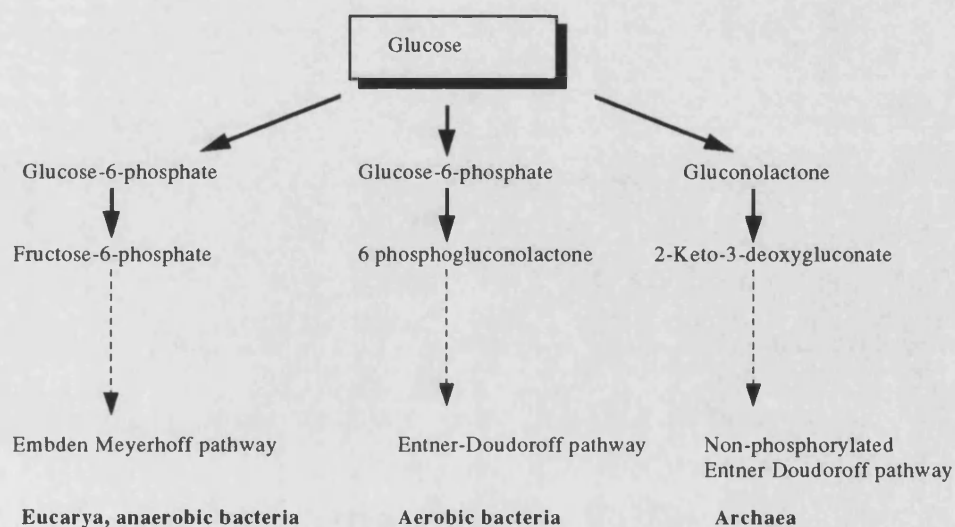


Fig. 1.3 Scheme depicting the differing pathways of glucose catabolism.

A number of glucose dehydrogenases, or enzymes closely related to GDH (in name at least) have been characterised (see table below).

Enzyme	Source	length (a.a) or size (kD)	Cofactor	oligomeric state	Role
GDH	<i>Tp. acidophilum</i> *	352	NADP/NAD	tetramer	Energy production
GDH	<i>B. megaterium</i>	262	NADP/NAD	tetramer	Sporogenesis
G6PDH	<i>L. mesenteroides</i> *	485	NADP	dimer	Pentose phosphate pathway
6PGDH	<i>Sheep</i> *	482	NADP	dimer	Pentose phosphate pathway
GDH	<i>Bovine liver</i>	59kD	NADP	tetramer	Endoplasmic reticulum catabolism
GDH	<i>Cyanobacterium</i>	100kD	NADP	monomer	Heterotrophic catabolism
GDH	<i>A. calcoalceticus</i>	478	PQQ	dimer	Energy production
GDH	<i>A. calcoalceticus</i>	801	PQQ	monomer	Energy production

Table 1. Various properties listed concerning characterised glucose (and glucose-like)

dehydrogenases. * denotes structure solved. GDH - glucose dehydrogenase (GDH), Bovine liver (Campbell *et al.*, 1982), *Cyanobacterium* (from a *Nostoc* sp.) (Juhász, *et al.*, 1986). G6PDH, glucose 6-phosphate dehydrogenase. 6PGDH, 6-phosphogluconate dehydrogenase. PQQ, pyrrolo-quinoline quinone.

Despite the commonality of name, and that GDHs generally appear to perform a catabolic role, there appears to be no further homology between these proteins. The monomeric and dimeric GDHs from *Acinetobacter calcoalceticus* (Schlunegger *et al.*, 1993) transfer electrons to ubiquinone and cytochrome b respectively, with the former being a membrane bound protein. The GDHs from *Bacillus* (Pal *et al.*, 1987) are thought to be homologous to the short-chain non-zinc requiring alcohol/polyol dehydrogenase family, of which the structure of one member, 2-hydroxysteroid dehydrogenase is known (Ghosh, *et al.*, 1994). The structures of glucose 6-phosphate

dehydrogenase from *Leuconostoc mesenteroides* (M. Adams, pers. comms.) and 6-phosphogluconate dehydrogenase from sheep have been determined (Adams *et al.*, 1991), but there is no evidence to suggest homology between GDH (from *Tp. acidophilum*) and these enzymes.

GDH from *Tp. acidophilum* has been isolated, purified and characterised ($K_m^{\text{NADP}} = 0.11 \text{ mM}$, $K_m^{\text{NAD}} > 30 \text{ mM}$, $K_m^{\text{Glucose}} = 10 \text{ mM}$) (Smith *et al.*, 1989). It was reported to be a dual cofactor specific enzyme, converting glucose to D-glucono-1,5-lactone in the first step of a modified Entner-Doudoroff pathway. GDH was subsequently cloned (construct:pTaGDH4), sequenced and expressed (using vector pMEX8 in *E. coli* TG1 cells). Gel filtration and sequence analysis revealed that GDH is a tetramer of 160 kD, with each monomer being composed of 352 amino acids (Bright *et al.*, 1993).

GDH from *Tp. acidophilum* shares no significant sequence identity to any known protein in the Swissprot data base, including the glucose dehydrogenases.

Nevertheless, the fingerprint motif GxGxxA was identified (192-197) which is present in some NADP binding proteins (see below). This information, in conjunction with the relative molecular mass of GDH tentatively assigned GDH to the long-chain Zn requiring ADH family (described later).

The nucleotide-binding domain

The nucleotide-binding domain (Rossmann fold) is a structurally conserved feature present in a number of dehydrogenases (Rossmann *et al.*, 1974, Ohlsson *et al.*, 1974, Grau, 1982), despite the corresponding lack of sequence identity in this region.

Classically, it is comprised of six parallel strands of a β -pleated sheet (β A to β F) and four helices (α B, α C, α E, α F) (fig 1.4), with the sheet exhibiting a marked 100° left

handed twist. The domain has an approximate 2 fold axis between βA and βD , with the first part of this domain forming the adenine binding pocket, and the second part forming the nicotinamide binding site.

The β -sheet (plus αB) is the most structurally conserved element of this domain, and represents the core of the domain, whereas the flanking helices (αC , αE , αF) of the

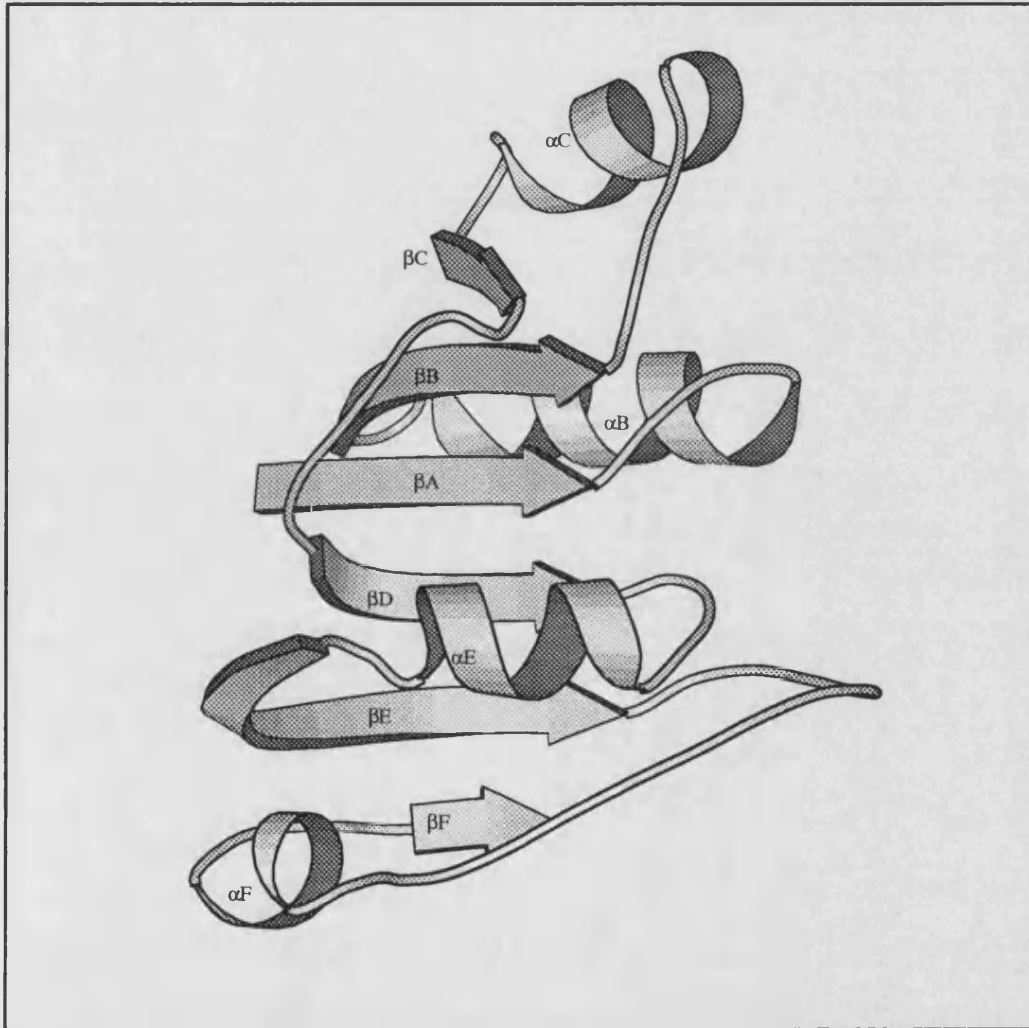


Fig 1.4. Schematic representation of the Nucleotide binding domain showing the major secondary structural elements. Produced using MOLSCRIPT(Kraulis, 1991).

dehydrogenases vary in length and orientation with respect to each other. In a number of dehydrogenases, $\beta A \alpha B \beta B$ forms a compact $\beta\alpha\beta$ unit (category 1), whereas some dehydrogenases, for example dihydrofolate reductase, possess a distorted $\beta\alpha\beta$ unit (category 2). The compact $\beta\alpha\beta$ unit is not an exclusive feature of the NAD(P) requiring enzymes, as it is found in many other enzymes, for example the adenylate kinases (Story & Steitz, 1992)

With more and more dehydrogenase structures being solved, the number of conserved structural elements within the classical Rossmann fold is slowly diminishing.

Glutathione reductase does not possess βE or βF (Karplus & Schulz, 1987), whilst glutamate dehydrogenase has an opposite orientation for the βC strand (Baker *et al.*, 1992)- which reduces the conserved elements to βA , βB , BD and αB . Possibly, some of these elements may lose their strictly conserved status when more dehydrogenase structures are solved.

Coenzyme specificity

The difference between NAD and NAD(P) is a phosphate moiety at the 2' hydroxyl of the adenine ribose. Some dehydrogenases are NAD specific, others are NADP specific, a few are true dual cofactor specific, whereas some dehydrogenases display a preferential binding for NAD/ NADP but can still bind NADP/NAD respectively, to a lesser degree (Tsai *et al.*, 1989). Much attention has been focused on the factors governing the differing coenzyme specificities exhibited by the dehydrogenases.

Rationalising the specificity from the structure is one approach. Changing the specificity, by rational site directed mutagenesis studies within the nucleotide binding domain, is another approach.

Despite the lack of sequence identity between the nucleotide binding domains, functional restraints must undoubtedly be applied to residues within this domain in order to bind the nucleotide. Wierenga *et al.* (1985) identified the helix within the $\beta\alpha\beta$ fold as playing a key role in the binding of the cofactor. The helix dipole of αB appears to be the main stabilising factor in the binding of the pyrophosphate moiety of NAD(P), and not the relatively weak hydrogen bonding interactions between the N-terminal backbone of αB and the oxygen's of the pyrophosphate moiety. Additionally, Wierenga *et al.* defined a fingerprint motif for the category I ADP binding $\beta\alpha\beta$ fold:

- Firstly, the GxGxxG sequence motif, which has served as a good diagnostic for NAD binding domains, occurs in the loop region between βA and αB . The first glycine allows a tight turn of the main chain, with special ϕ , ψ angles. Any other residue apart from the second glycine would sterically hinder the binding of the nucleotide. The third glycine is important to provide space for a close interaction between the β strands and the α helix.
- Secondly, six relatively small hydrophobic residues, indicated in the sequence alignment (fig. 8.1), form the hydrophobic core of the $\beta\alpha\beta$ unit.
- Finally, a conserved negatively charged residue, present at the C-terminal end of βB , which forms hydrogen bonds to the 2' hydroxyl group of the adenine ribose moiety.

This consensus sequence pattern of residues has enabled the recognition of NAD binding sites within the protein database using a property pattern approach (Bork & Grunwald, 1990).

In regards to NADP binding proteins, the sequence motif GxGxxA has been identified (Hanukoglu & Gutfinger, 1989). The alanine is believed to prevent a close interaction between the β strands and the α helix, which causes a change in the coenzyme specificity (Scrutton *et al.*, 1990). This sequence motif, in conjunction with a positively charged residue at the C-terminal end of β B is thought to confer NADP specificity.

Numerous protein engineering experiments have attempted to change the coenzyme specificity of dehydrogenases. However, factors governing coenzyme specificity are not as clear cut as previously thought. Mutagenesis studies on glyceraldehyde 3-phosphate dehydrogenase from *Bacillus stearthermophilus* focused on changing the specificity from NAD to NADP (Clermont *et al.*, 1993). Clermont *et al.* concluded that the presence of a positively or negatively charged residue at the end of β B is an important factor governing the specificity of NADP or NAD. However, additional factors such as steric and conformational constraints are involved. The presence of Asp (or Glu) at the end of β B, the absence of a hydrophilic adenosine subsite, and the lack of space are believed to be the reasons why NAD proteins cannot bind NADP. Scrutton *et al.*, altered the specificity of *E.coli* glutathione reductase from NADP towards NAD. Mutating the two arginines (at the end of β B) to Met and Leu, the alanine (of GxGxxA) to glycine conferred NAD activity. However, a suite of further mutations in the $\beta\alpha\beta$ fold were necessary to markedly change the coenzyme specificity of the enzyme. Although, in these types of protein engineering experiments, the mutants were not as catalytically efficient as the wild type enzyme, indicating that specificity-determining residues cannot be the only consideration in redesigning the coenzyme specificity of an enzyme.

In addition factors governing NAD or NADP specificity are not global, as exemplified by the glutamate dehydrogenase family (GLUTDH) (Baker *et al.*, 1992). The GLUTDH family contain enzymes that are NAD specific, NADP specific and dual cofactor specific. The GLUTDH from *C. Symbiosum* (structure solved, Baker *et al.*, 1992) is NAD specific, yet it possesses the GxGxxA motif, and has a glycine residue corresponding to the Asp residue at the end of β B. Furthermore, some dual cofactor and NADP dependent GLUTDHs contain an Asp or Glu at this site. Baker *et al.* (1992) identified a direct/indirect hydrogen bonding scheme between the adenosine ribose and the GxGxxA motif and GxGxxG motif respectively, with both types of interaction being independent of coenzyme specificity. Baker *et al* also envisaged a protonated form of the 2' phosphate hydrogen bonding to the Asp/Glu of the NADP-utilising GLUTDHs.

Mode and action of nucleotide binding

Structural superposition of the nucleotide binding domain have led to the finding that the mode of nucleotide binding is also well conserved (Wierenga *et al.*, 1985). The NAD(P) binds in an extended conformation to the C-terminal ends of the β -strands of the Rossmann fold. The adenosine moiety sits in a hydrophobic pocket provided by the $\beta\alpha\beta$ fold. The positioning of the adenosine moiety is structurally conserved, whereas there is more structural deviance in the site of the nicotinamide ring.

Dehydrogenases accept protons directly from the substrate in a stereospecific manner. Based on this stereospecificity, dehydrogenases have been divided into two classes, A and B. Alcohol dehydrogenase is an example of an A-side specific enzyme where the pro-R hydrogen of the alcohol is transferred to the A-side of the nicotinamide ring.

Alcohol dehydrogenase

The enzyme alcohol dehydrogenase has been characterised from a whole host of organisms, ranging from the Bacteria, Eucarya and the Archaea. The advent of molecular biology has resulted in a spiralling number of ADH genes being cloned and sequenced. The sequence information from these organisms have enabled the ADH superfamily to be split into 3 main families (Jörnvall *et al.*, 1987): the non-metal requiring, the iron containing and the long-chain zinc-requiring enzymes (Zn-ADH). (fig. 1.5). Of the Zn-ADHs, only two structures have been reported - that of horse liver alcohol dehydrogenase (LADH) (and more recently human alcohol dehydrogenase (HUD) (Hurley *et al.*, 1994). The LADH has been extensively studied (see below), and the structure has aided in the analysis of structure-function relationships of other members of the Zn-ADH family, as well as a phylogenetic analysis between the members (Sun & Plapp, 1992). Sequence analysis has also revealed that sorbitol dehydrogenase (Eklund *et al.*, 1985), threonine dehydrogenase (Aronson & Somerville, 1989) and ζ -crystallin (Borrás *et al.*, 1989) are related to the Zn-ADH family. The Zn-ADH family are generally 350 - 380 amino acids long, possess a centrally located nucleotide binding domain, as judged by the GxGxxG/A sequence motif, and require zinc for catalytic activity. The majority of these enzymes are NAD dependent, whereas only a few of these enzymes, notably the bacterial Zn-ADHs exhibit NADP specificity. The Zn-ADHs are either dimeric or tetrameric. Higher order Eucarya (plants, animals) are typically dimeric, whereas yeast (YADH), Bacterial, Archaeal and sorbitol dehydrogenases (SDH) are usually tetrameric.

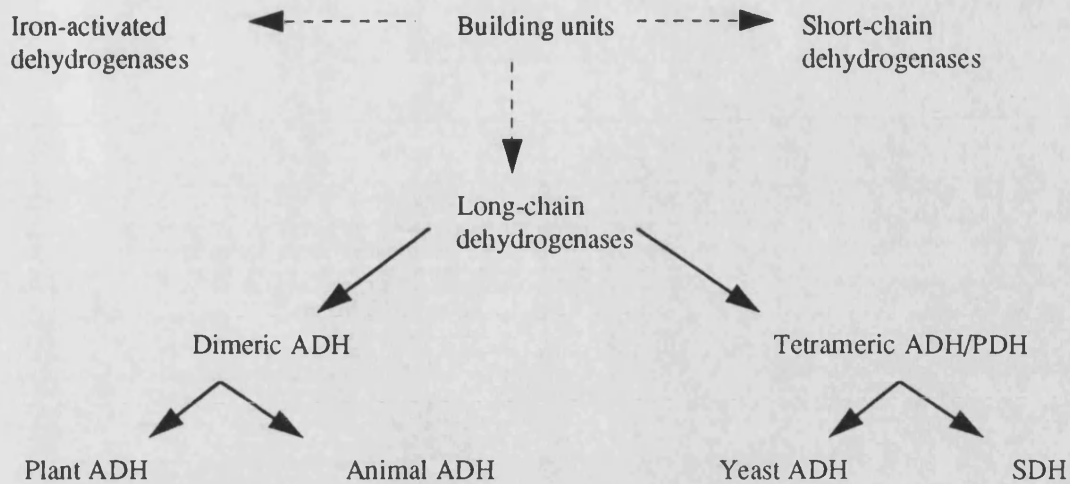


Fig. 1.5 Schematic, simplified representation of the alcohol dehydrogenase family (derived from Jörnvall *et al.*, 1987).

Two multiple sequence alignments of the Zn-ADH family have been reported. Jörnvall *et al.* (1987) compared 17 different members (11 different species) using pairwise alignment, which was based on the secondary structural elements of LADH. Later, with the arrival of more sequences, Sun & Plapp (1992) compared 47 members (31 different species) using a progressive sequence alignment (see chapter 7 for discussion).

Jörnvall found that there were 22 strictly conserved¹ residues throughout the family, whereas the more comprehensive study by Sun & Plapp revealed that there were only 9 residues (eight gly :66, 71, 77, 86, 192, 201, 204, 236 one val 80) strictly conserved throughout the Zn-ADH family. The number of strictly conserved residues throughout the family are destined to decline when more Zn-ADH sequences are compared.

¹ within a given alignment, a residue is deemed strictly conserved if all the residues are identical at a specified position.

The sequence identity between members of the Zn-ADH family is generally low, with the plant Zn-ADHs being more conserved than the animal ones. This lack of strictly conserved residues and sequence identity probably reflects the generally broad substrate specificity and ambiguous physiological role exhibited by the Zn-ADHs, although specific physiological functions have been identified in some Zn-ADH members.

In addition to the catalytic zinc, the dimeric and tetrameric yeast Zn-ADHs possess a structural zinc ion/monomer. The function of this structural zinc is largely unclear (however see chapter 7), nevertheless a recent study has shown that it plays an important role in maintaining the conformational stability of a tetrameric yeast Zn-ADH (Magonet *et al.*, 1992).

Horse liver alcohol dehydrogenase (and the models)

LADH, a dimeric enzyme of 80kD, composed of identical subunits (374 amino acids per monomer), is an NAD dependent enzyme that exhibits a broad substrate specificity towards primary alcohols and also oxidises some secondary alcohols (Sund & Theorell, 1963). The structure and mechanism of action of LADH has been extensively studied (Klinman, 1981; Petterson, 1987 and references therein).

The structure of the apo-form has been determined to 2.4Å resolution (Eklund *et al.*, 1976). The monomer can be split into two domains, an N- and C-terminal catalytic domain (1-175, 319-374 respectively) and a centrally located nucleotide binding domain, which are separated by a deep active site cleft (fig. 1.6).

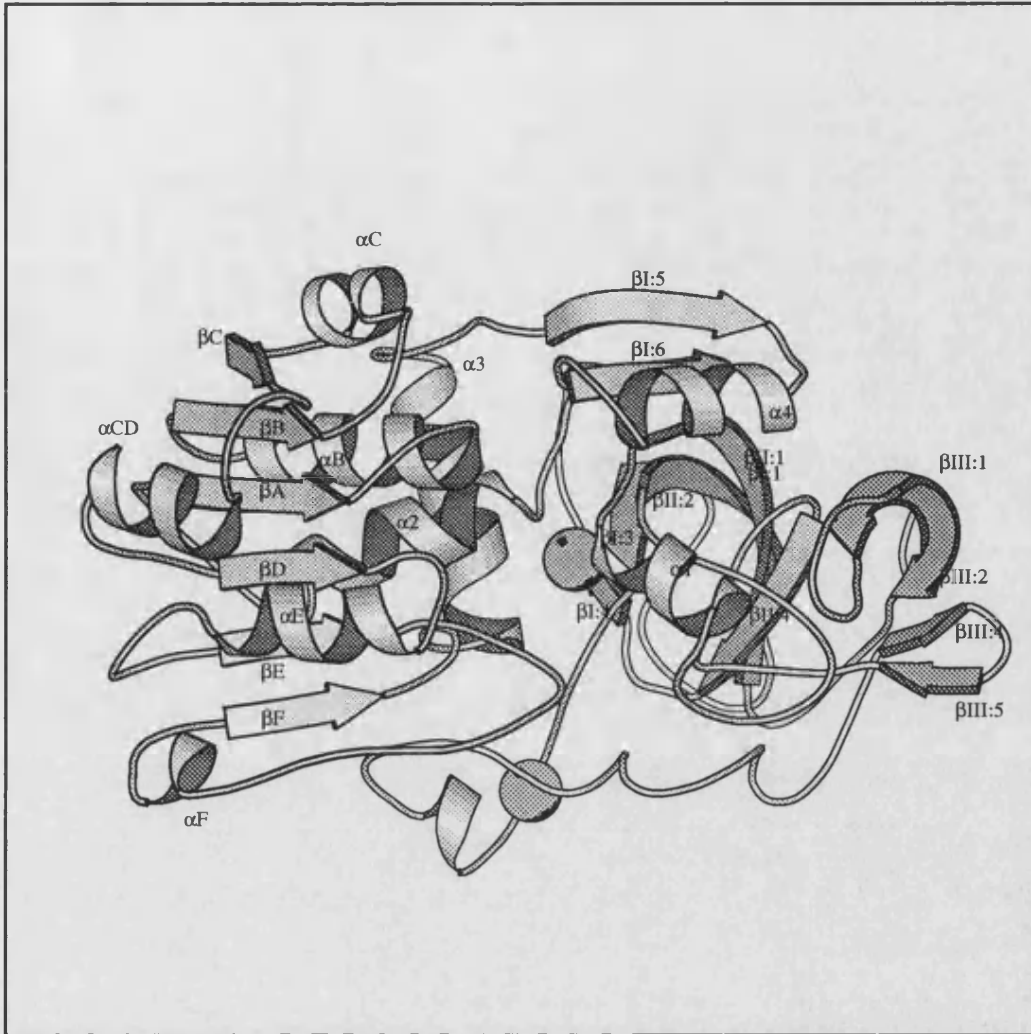


Fig. 1.6. Schematic representation of the LADH monomer, produced using MOLSCRIPT (using DSSP to define the secondary structural elements). The secondary structural elements defined by Eklund *et al.* are: $\beta III:1$, 9-14; $\beta III:2$, 22-29; $\beta II:1$, 33-40; $\beta I:1$, 41-45; $\alpha 1$, 46-53; $\beta III:3$, 62-65; $\beta I:2$, 68-71; $\beta II:2$, 72-78; $\beta II:3$, 86-87; $\beta I:3$, 88-92, $\beta III:4$, 129-132, $\beta III:5$, 135-138; $\beta III:6$, 145-146; $\beta II:4$, 148-152; $\beta I:4$, 156-160; $\alpha 2$, 168-179; αA , 180-188; βA , 193-199; αB , 201-215; βB , 218-223; αC , 229-236; βC , 238-243; αCD , 250-258; βD , 263-269; αE , 271-280; βE , 287-293; βS , 299-303; $3^{10}S$, 304-311; βF , 312-318; $\alpha 3$, 324-338; $\beta I:5$, 347-352; $\alpha 4$, 355-365; $\beta I:6$, 369-374.

LADH possesses the conserved structural features of the Rossmann fold, but also contains an additional helix (αCD) and a strand (βS) - of which βS participates in

subunit-subunit interactions. The domain possesses the GxGxxG motif and an Asp 223 which binds the 2'OH of the adenosine ribose, and thought to confer specificity towards NAD. Two interconnecting helices ($\alpha 2$ and $\alpha 3$) connect the catalytic domain which is comprised mainly of 3 anti-parallel β -pleated sheets (βI , βII , βIII) plus 4 helices ($\alpha 1$ - $\alpha 4$). The two zinc ions are found within the catalytic domain with the catalytic and structural zinc at the bottom and mouth of the active site cleft respectively. The ligands to the catalytic zinc, Cys 46, His 67 and Cys 174, are well conserved throughout the Zn-ADH family. In the apo-form a water molecule coordinates to the zinc to complete a tetrahedral arrangement around the zinc. The catalytic zinc ion is directly involved in the catalytic mechanism of LADH (Eklund *et al.*, 1982)(see below).

The structural zinc is contained within a lobe at the surface of the enzyme, and obscures the active site cleft. The four cysteine ligands (97, 100, 103, 111) are well conserved throughout the structural zinc - containing Zn-ADHs. As stated above, no definitive role has been identified for the structural zinc (especially in the dimeric enzymes).

LADH dimerises mainly via the carboxyl ends (βF and βS) of the nucleotide binding domain, forming a 12 stranded β -sheet with the 2 pairs of the parallel sheets being joined in an anti-parallel fashion. The subunit-subunit interactions are generally hydrophobic, which form the central core of the molecule, with the catalytic domains at each end of the molecule. A minor site of interaction is via the structural lobe and a loop between the αE and βE of the nucleotide-binding domain.

From the LADH structure, monomeric models of YADH (Jörnvall *et al.*, 1978) and SDH (Eklund *et al.*, 1985) which are tetrameric enzymes, have been proposed. One

notable difference between the tetrameric enzymes and the dimeric enzymes is that the dimers possess a stretch of 21 amino acids (119-139) that forms a surface loop (115-140), which based on the sequence alignment, is not present in the tetramers. In SDH, Glu 174 was aligned to be structurally equivalent to Cys 174 of LADH, and believed to act as a ligand to the catalytic zinc. This Cys-Glu mutation, in conjunction with a Phe 93 - Val 93 mutation was correlated with the ability of SDH to oxidise the polyol, sorbitol.

In addition, conjectures regarding the tetrameric nature of these monomeric models were made. The main area of putative structural differences between LADH and the models (longer loop, residues 55-60; lack of surface loop, residues 119-139; shorter region, residues 305-315 (LADH numbering)) lie on one face of the enzyme, and were hypothesised as acting as areas of subunit-subunit interactions different from that found in the dimer. However, no completely plausible tetrameric model could be established.

The mode of binding of inhibitors and coenzyme analogues, that do not change the conformation of LADH, have also been structurally determined (Eklund & Branden, 1979, and references therein). One of these coenzyme analogues, ADP-ribose, which represents the adenosine-ribose moiety of NAD, was shown to be positioned in a structurally equivalent manner to that of the same moiety of other NAD-dehydrogenase complexes (Eklund *et al.*, 1984).

NAD is known to change the conformation of LADH. Ternary complexes of reduced coenzyme/DMSO (Eklund *et al.*, 1981) and NAD/p-bromobenzyl alcohol (Eklund *et al.*, 1982), and more recently NAD/2,3,4,5,6-pentafluorobenzyl alcohol have been

structurally determined (Ramaswamy *et al.*, 1994). These holo-enzyme structures represent the 'closed' form of the enzyme, from which the following have been reported:

- structural differences between the apo- and holo- forms (Eklund & Brändén, 1979).
- The mechanism and consequences of domain closure (Colonna-Cesari *et al.*, 1986)
- Substrate/coenzyme/enzyme interactions (Ramaswamy *et al.*, 1994).

The conformational change is essentially a rigid body rotation, with the catalytic domain undergoing a 10° rotation with respect to the nucleotide-binding domain, resulting in the catalytic domain moving closer to the nucleotide-binding domain.

The catalytic domain moves closer to the coenzyme-binding domain by a sliding motion as opposed to a least-motion path of cleft closure. Two glycine rich regions at either end of the nucleotide binding domain (173-175, 318-321), were deduced to act as hinge regions for this domain motion (Colonna-Cesari *et al.*, 1986).

Only a few changes in loop conformations (47-53 and 292-300) are necessary to accommodate this conformational change, which brings no new residues into the active site but serves to narrow the active site cleft by 1-2Å. Weak substrate binding to the apo-form LADH is observed as no new residues are brought into the active site, however productive binding of the substrate to the enzyme is only possible after the conformational change, which correlates with the ordered mechanism of LADH (Theorell & Chance, 1951). This conformational change has the effect of shielding the catalytic zinc from the solution, and removing the water that is ligated to the zinc and Ser 48, so that hydride transfer takes place in a water free environment. The alcohol group is believed to directly interact with the zinc ion, which would depress the pKa of the alcohol group and enable the alkoxide to be readily formed. An extensive

hydrogen bonding network is established upon binding of substrate and coenzyme (Ramaswamy *et al.*, 1994), which is believed to promote the proton dissociation from the alcohol and relay the proton from the hydrophobic environment to the aqueous phase (fig. 1.7). From site-directed mutagenesis studies, His 51 is known to confer an important role in the proton relay system (Ehrig *et al.*, 1991).

The pro-R hydrogen is positioned to facilitate rapid hydride transfer to the C4 atom of the A-side of the nicotinamide ring. The aldehyde rapidly dissociates from the complex, whereas dissociation of the reduced coenzyme is rate limiting.

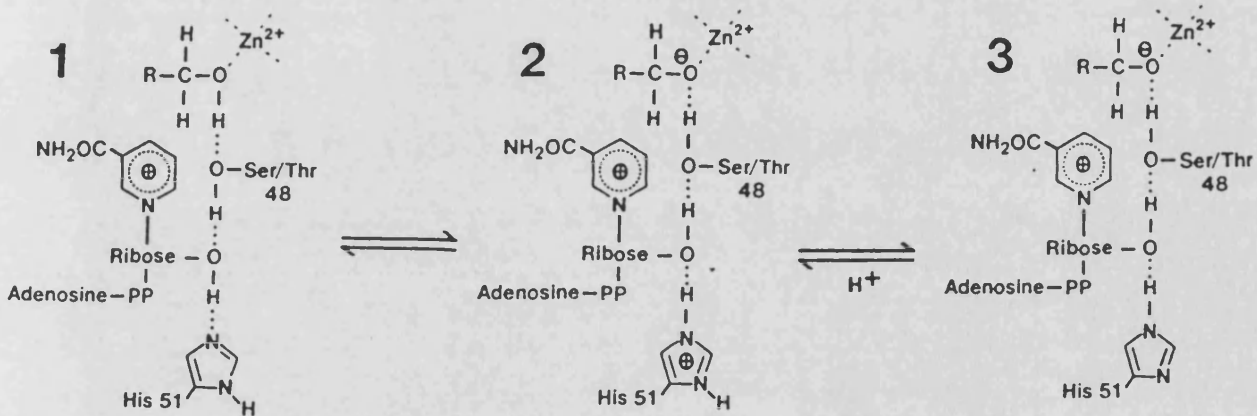


Fig. 1.7. Scheme for hydrogen bonding and ionisation of the proton-relay system.

Aims of project

Quite simply, the aim of the project was to solve the structure of the Archaeal enzyme, glucose dehydrogenase. Structural studies on Archaeal enzymes may reveal evolutionary relationships between proteins, which primary sequence data alone may not disclose, and may give an insight into how proteins are adapted to extreme environmental conditions. In addition, the structure may aid to rationalise the basis of the observed coenzyme and substrate specificity.

Chapter 2

PURIFICATION AND CRYSTALLISATION

Introduction

Purification with the ultimate goal of crystallisation requires the protein sample to be as homogeneous as possible (Abergel *et al.*, 1991). Glucose dehydrogenase from *Tp. acidophilum* has been cloned, sequenced, expressed and subsequently purified (Bright *et al.*, 1993). The salient step of the reported purification scheme (Fig.2.1, scheme 1) was that it capitalised on the thermostable properties of the glucose dehydrogenase via the incorporation of a heat treatment step (Bright, 1991).

Proteins are complex macromolecular systems and diverse in nature - consequently there are no empirical formulae for crystallising proteins (McPherson, 1990). To maximise the chance of obtaining crystals, as many conditions as possible should be searched. However, these conditions need to be screened as efficiently as possible, as a crystallisation screen is a time consuming process, plus the supply of protein may be limited. The Incomplete Factorial approach (INFAC)(Carter *et al.*, 1988) and the 'Magic 50' (Jancarik & Kim, 1991) represent two such efficient screens.

The INFAC approach is a crystallisation matrix that encompasses all factors deemed potentially important in the crystallisation of a protein, such as a range of precipitants, cations, anions, cofactors, pH, temperature and protein concentration. These factors are randomised to produce a unbiased crystallisation matrix. A scoring system for crystal quality for each condition is established and a subsequent statistical analysis (multiple regression) of the results enables a rapid determination of the factors that have a positive effect on crystallisation (Carter *et al.*, 1988), and provides a starting point for a finer crystallisation grid.

The 'Magic 50' is comprised of 50 crystallisation conditions; it is based on the INFAC design, but the conditions used are biased towards previously published crystallisation conditions. This method has proven to be an effective method for obtaining crystals (Jancarik & Kim, 1991).

Should a protein prove difficult to crystallise, then attempts should be made to remove any impurities and in addition to reduce any source of microheterogeneity as: molecular isoforms, variations arising from post-translational modification and modification of the protein during the course of purification. If a protein fails to crystallise following these 'clean-up' operations, then the protein may be inherently too flexible or unstable to crystallise. Limited proteolysis (McPherson & Spencer, 1975) and chemical modification (Abeyasinghe *et al.*, 1991; Stillmann *et al.*, 1992) are two such approaches that have reduced the inherent flexibility/instability of a protein and aided in subsequent crystallisation.

Once the initial condition(s) at which crystallisation of a protein has been achieved, the quality and size of the crystal should be improved (if needed). Crystal seeding, both micro- and macroseeding, are well documented techniques that have led to larger crystals (Thaller et al., 1981,1985; Stura & Wilson, 1990).

Bigger is better

For a given protein, the intensity of the diffraction pattern is directly proportional to the volume of the crystal. Thus the larger the crystal the stronger the diffraction, often leading to diffraction to higher resolution. Efforts should therefore be made to maximise the size of the crystal. Crystal formation can be divided into two stages:

- Spontaneous nucleation of the crystal seed.
- Subsequent crystal growth.

The aim of seeding is to achieve crystal growth only. Generally this is accomplished by using a precipitant concentration slightly lower than that in which crystals appear (see Appendix 1, Crystallisations, Fig. 1). Attention must be focused on when to transfer the crystal seed from the original solution to the newly established condition, *i.e.* the optimum pre-equilibration time must be determined for seeding. Another facet of macroseeding is that the seed must be washed during transfer using appropriate solutions. These solutions serve to remove any satellite crystals attached to the seed, and to slightly dissolve the surface of the crystal to produce a fresh surface for continued crystal growth.

Results

The purification scheme reported (scheme 1) produced pure protein as judged by SDS Phast Gels (8-25%)(Appendix 1, Protein Methods), although IEF Phast Gel analysis revealed microheterogeneity within the sample (R. Mackness, 1991). This microheterogeneity was believed to have arisen from the heat treatment step. Thus from reasons outlined in the introduction, this step was replaced by a milder solvent extraction step (scheme 2, Fig. 2.1) reported by J. Bright (PhD. Thesis 1991). The resultant purification scheme did not significantly reduce the microheterogeneity of the purified sample, hence the heat treatment step was reinstated as it was deemed more efficient.

Despite the microheterogeneity, crystallisation trials based on the previously reported conditions (Bright *et al.*, 1991) were constructed. Crystal formation was not reproducible, therefore the crystallisation search was extended by:

- Widening the PEG (polyethylene glycol) matrix around the reported conditions.
- The 'Magic 50' screen.
- The INFAC approach.

No crystals suitable for diffraction analysis were obtained, but the INFAC approach determined that PEG was a positive crystallisation factor, whereas low temperatures and high protein concentrations had detrimental effects on crystal formation. The lack of crystals implied a possible problem with the protein sample.

Scheme 1: Heat treatment → Gel filtration → Anion exchange → Gel filtration

Scheme 2: Solvent extraction → Gel filtration → Anion exchange → Gel filtration

Scheme 3: Heat treatment → Dye Ligand → Anion exchange → Gel filtration

Scheme 4: Heat treatment → Solvent extraction → Gel filtration → Anion
exchange → Gel filtration

Fig. 2.1. Range of purification schemes explored. For more details of the methods see Appendix 1, Protein methods.

Subsequent analysis of the purified protein concentrate (15mg/ml) on an SDS PAGE gel (10%) revealed a multitude of impurities (Fig. 2.2a). This was believed to be a major contributory factor in the failure to produce crystals, hence steps were undertaken to re-purify the protein.

A range of purification protocols (schemes 3 and 4, Fig. 2.1) incorporating Dye ligand chromatography (Appendix 1, Protein methods), heat treatment, solvent extraction and gel filtration were explored. The Dye ligand chromatography step yielded essentially pure protein (Fig 2.2b), however yields from this column were low (1-2 mg/l of culture). Nevertheless, using scheme 3, sufficient protein was recovered to establish a crystallisation trial (the 'Magic 50'), from which small, regular crystals were formed under the conditions :-

Precipitant: 20% PEG4K, 10% propan-2-ol; **Buffer:** 0.1M Hepes (pH 7.5);

Protein concentration: 6-8mg/ml; **Temperature :** 25 °C.

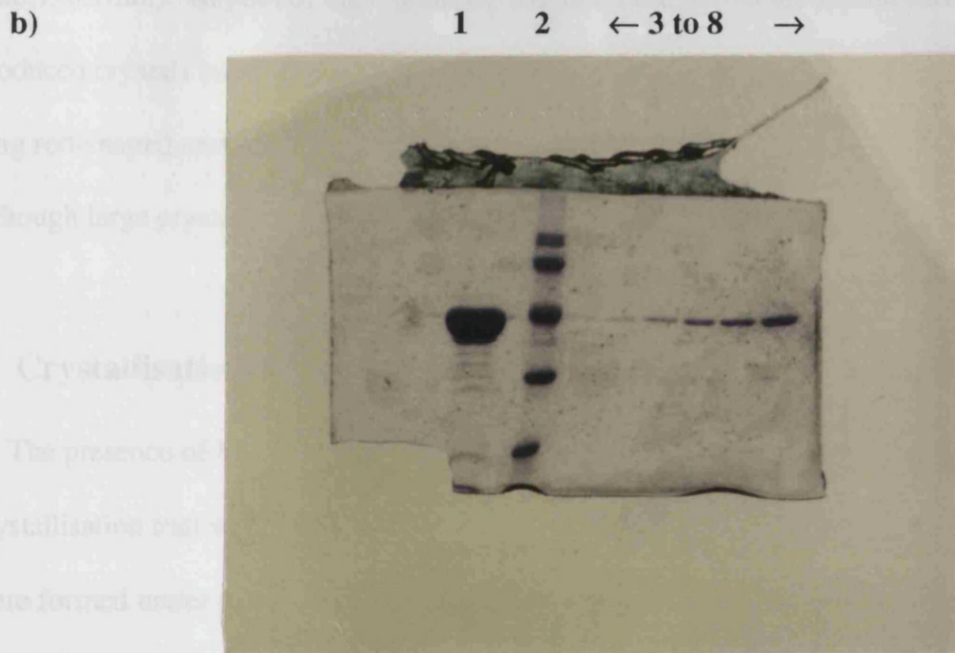
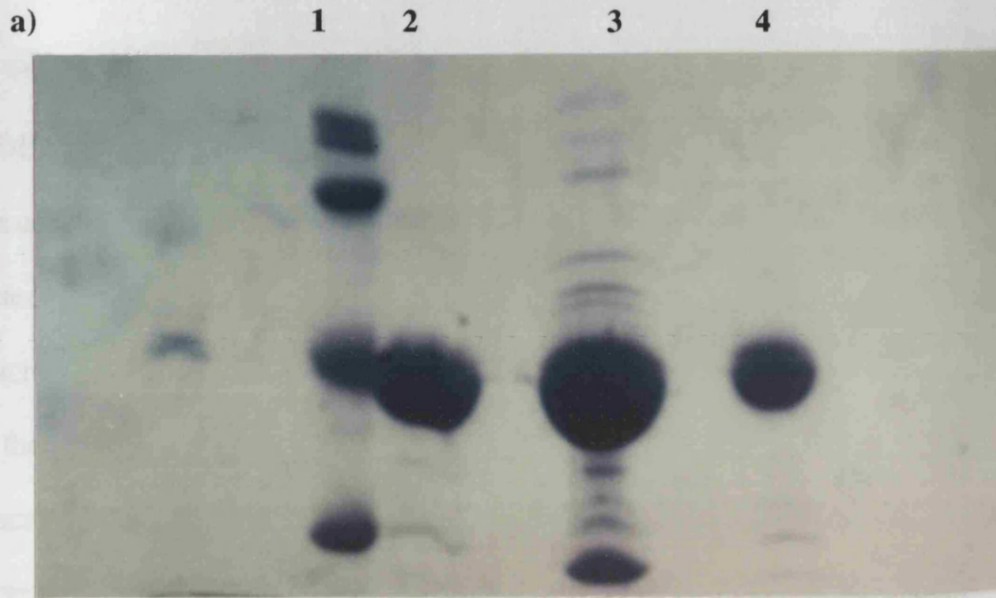


Fig. 2.2 a) SDS-PAGE gel displaying impurities of the 'pure' GDH concentrate. **Lanes:** 1: Markers; 2 & 4: re-purified GDH sample; 3: the 'pure' GDH concentrate **b)** Pure GDH fractions from the Dye ligand chromatography purification step. **Lanes:** 1: GDH concentrate; 2: Markers; 3-8: Dye ligand eluent. Molecular weight of markers are: 97.4, 66.2, 45, 31, 21.5kD (descending order).

Purification scheme 4, which provided a much greater yield of protein (5-8mg of GDH/l of culture), also produced small crystals under the same conditions. However, the crystals were too small (0.5x0.05x0.05mm) for in-house X-ray data analysis. Attempts were made to increase the size of the crystal by various seeding techniques. Microseeding demonstrated the potential to increase the size of the crystal, however due to the inability to reliably produce large single crystals, attention was focused on macroseeding. The optimum conditions for macroseeding were determined (Appendix 1, Crystallisations, Fig. 1) and led to a significant increase in the size of the crystal (1x0.1x0.1mm). Repetitive macroseeding increased the size of the crystal further, which produced crystals suitable for X-ray diffraction studies (Fig. 2.3). The crystals were of long rod-shaped morphology; the average dimension of the crystal was 1.5x0.2x0.15mm, although large crystals (1.5*0.3*0.2mm) were occasionally obtained.

Crystallisation with Cofactor

The presence of NADP completely disordered and dissolved the crystals, hence a crystallisation trial was briefly experimented with (the 'magic 50'). Sheets and needles were formed under some of the conditions (results not shown). From kinetic studies (see Chapter 10), and in hindsight, crystallisation in the presence of cofactor alone would lead to a heterogeneous population of GDH species, and thus present a difficult sample to crystallise. Crystallisation of GDH in the presence of cofactor and an inhibitor, or an abortive ternary complex, which would lock GDH in the 'closed form', should be a more fruitful approach.

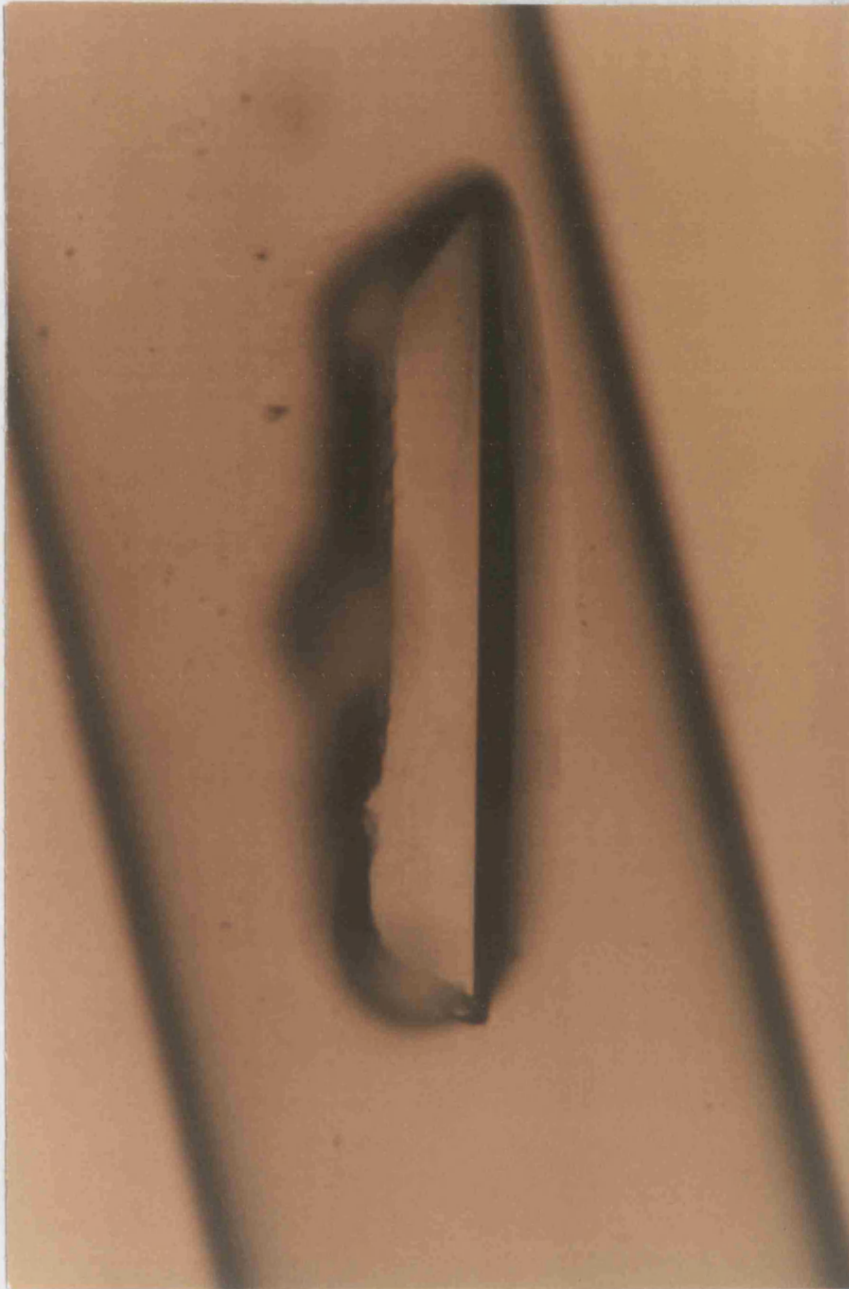


Fig. 2.3 A typical GDH crystal.

Chapter 3

NATIVE DATA COLLECTION AND ANALYSIS

Introduction

Once suitable sized crystals are obtained, the next progression is X-ray data collection. The theory of X-ray crystallography is discussed briefly in Appendix 2. The validity of a crystalline protein sample in representing the true structure of the protein in solution is frequently raised by non-crystallographers. The demonstration of good correlation between crystal structures and NMR solved structures; the phenomenon of catalytically active protein crystals (Rhodes, 1993; Blundell & Johnson, 1976); the solvent content of crystals being typically 50% (Matthews, 1968); different crystal forms possessing very similar structures, and the observation that a 3-D crystal structure permits a rational and effective approach to inhibitor/drug design (Verlinde & Hol, 1994) should however dispel any doubts on the validity of X-ray crystallography. X-ray diffraction data on native and GDH derivative crystals (see chapter 5), were collected on the in-house Siemens area detector, and higher resolution native data using synchrotron radiation at station 9.5, SERC Daresbury, Laboratories, U.K. The methods employed for data collection and reduction were different for each source.

In-house detector

The crystals, bathed in a minimal amount of mother liquor, were mounted in a small quartz capillary tube, and the capillary sealed with beeswax (Rayment, 1985). The sample was positioned on the goniometer head with the translational sledges enabling centering of the crystal in the X-ray beam. The crystal can be rotated 360° about the ϕ axis, and $>180^\circ$ about ω . The 3-axis goniostat has κ fixed at 45° . The X-rays are generated by a rotating anode X-ray source, which was routinely operated at 35-40 kV, 70-80 mA. The chamber of the rotating anode generator is maintained under constant vacuum, and water is continuously circulated around the copper anode to dissipate the heat generated. Monochromatic Cu-K α radiation is obtained from the Cu X-ray spectrum (wavelength, $\lambda = 1.542\text{\AA}$) using a graphite monochromator: the resultant beam is passed through a collimator, typically 0.3mm in diameter.

The diffracted X-ray beams were collected on a Siemens multiwire area detector (model:X-100), which has a concave beryllium window facing the X-ray source and a xenon/carbon dioxide gas filled chamber. Within the detector there are 3 parallel planar electrodes: two cathodes sandwiching an anode. One cathode is the beryllium window, while the other cathode and anode are composed of 256 wires which are perpendicular to each other: which partitions the detector window into a 512 x 512 array of pixels. When an X-ray photon passes through the beryllium window, it is absorbed by Xenon, with a concomitant ionisation of that molecule. The free electrons produced move towards the anode, inducing an electrical signal, which is decoded by the detector to give an x and y co-ordinate for the event.

The crystal to detector distance is of paramount importance in the data collection strategy, and is governed by the unit cell dimensions of the crystal. The operational

parameter, empirically about one tenth of the largest unit cell dimension, determines the crystal to detector distance (cm) (G. Taylor, pers. comms.), and allows for optimal spatial resolution of the detected spots. The detector is mounted on a horizontal arm, where θ_c is the angle formed by the X-ray beam and the normal to the centre of the detector. The crystal to detector distance, radius of the detector, and the θ_c swing of the detector determines the maximum and minimum resolution of the data to be collected:

$$d_{\max} = \frac{\lambda}{2 \sin \theta_{\max}}$$

$$2\theta_{\max} = \theta_c + \tan^{-1}\left(\frac{r}{D}\right)$$

d_{\max} = maximum resolution, D = crystal ↔ detector distance, r = radius of detector,

θ_c = detector swing, θ_{\max} = angle of maximum resolution, λ = wavelength of X-ray beam.

Hence, the greater the detector swing, the greater the resolution at the edge of the detector.

Prior to data collection, a flood field and brass plate image were collected, using an ^{55}Fe emitter, with the detector bias adjusted accordingly, at the determined crystal ↔ detector distance. The flood field is used to compensate for local heterogeneity at the detector surface. The radial distortion of the detector is determined by measuring the pattern cast by a precision brass plate with holes machined on a 2-D grid, which is then used to correct each image collected. The brass plate is transferred to the MicroVAX and used by XDS. The time of exposure depends on detector distance. For GDH, exposure times of 10 minutes for the floodfield and one hour for the brass plate were typical.

Data collection and analysis (Area detector)

Contiguous non-overlapping frames were collected, with 10 oscillations of 0.25° about the ω axis/frame. The initial quality of the diffraction obtained and subsequent radiation damage was visually assessed using SADIEN (Siemens Area Detector Integrated Environment program). If the quality of the diffraction pattern deteriorated notably during the course of data collection, the crystal was translated along its main axis and data collection was resumed on an unexposed area of the crystal. All data collected were analysed using XDS (Kabsch, 1988) (see fig. 3.1). The program consists of several steps:

XYCORR: Creates a look up table (XYCORR.TABEL) which accounts for the positional distortion of the detector images (using the brass plate as input). **INIT:** Calculates an initial background at each detector pixel (BKGPIX.XDS) from the first 30 frames collected, dividing the detector into 9 areas, and decides what constitutes the active area of the images (MODPIX.XDS) **COLSPOT:** The centroids of strong spots are located over a given number of images which are significantly above using the calculated background value (more than 200 spots should be found for auto-indexing (SPOT.XDS)). **IDXREF:** autoindexes the spots found by COLSPOT, and refines the unit cell dimensions and the crystal orientation. The direct beam position is refined, while the crystal to detector distance is fixed (XPARM.XDS). **IDXREF** predicts where reflections should occur, and indexes these spots. Typically, a standard deviation of spot position < 1 pixel, and a standard deviation of spindle position $< 0.12^\circ$, are acceptable data processing errors. **COLPROF:** Collects the spot profiles from all the data frames. Every 15 frames, the crystal orientation and unit cell are

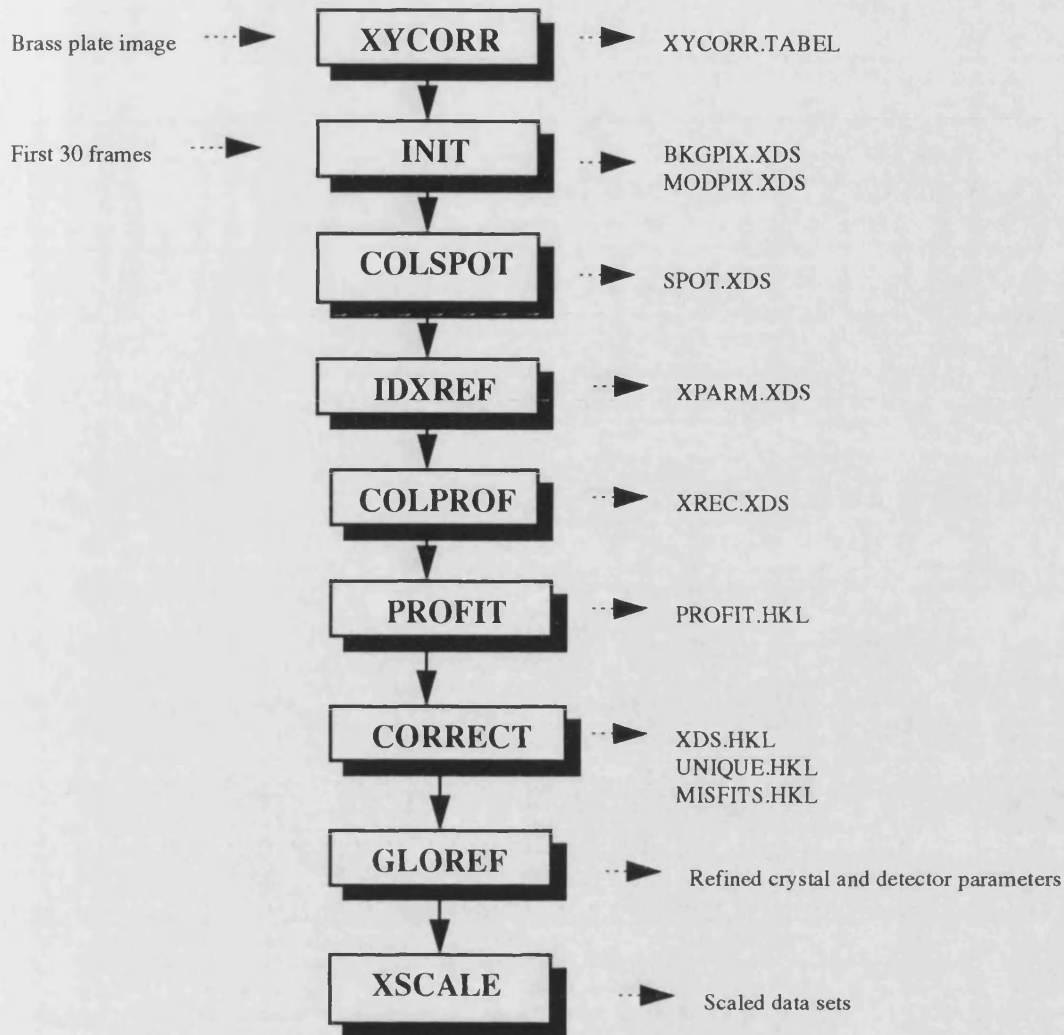


Fig. 3.1 A flowchart depicting the stages of the data reduction process using XDS.

refined using all reflections collected in the previous 15 frames. The strong reflections are superposed and a normalised 3-D reflection profile is represented by nine 9×9 matrices. A compact centered intensity profile indicates good data collection, whereas streaky and diffuse profiles are indicative of crystal slippage or radiation damage.

PROFIT: The detector face is divided into 9 areas: a bulls eye and 8 radiating wedges with each area having its own 3-D profile. The program integrates the pixel values within the profile, estimates the background and calculates an intensity and standard

deviation for each reflection. **CORRECT**: Corrects the integrated intensity of each reflection for crystal decay and merges symmetry related reflections together.

GLOREF: Post-refinement of the unit cell and camera parameters using all the data.

XSCALE: Applies a semi-empirical absorption correction and can scale multiple data sets.

X-ray synchrotron

In this device, a beam of electrons accelerated to 0.999x velocity of light is constrained by a magnetic field to travel through a circular path (under vacuum). As the path of the electrons is bent by a magnet, electromagnetic radiation is emitted tangentially. The radiation spectrum is continuous between the X-ray and infrared region, and current synchrotron devices are designed with the insertion devices such as wigglers to produce intense X-rays to shorter wavelengths.

The advantages of the synchrotron source when compared to the in-house source are:

- The high intensity of the X-ray beam allows faster data collection and collection of higher resolution data.
- A wide X-ray spectrum is emitted, unlike the fixed wavelength of the copper anode, which enables the wavelength at the station to be selected - which may prove useful in trying to solve the phase problem.
- The shorter wavelengths which can be used at the synchrotron to reduce absorption and radiation damage.

The detector at Station 9.5, Daresbury, is an 18cm Mar research imaging plate. The imaging plate has a phosphor screen which temporarily stores the diffracted X-ray image, which is subsequently read by a scanner. At station 9.5, this scanning

procedure results in a 3 minute off-time. The high sensitivity of image plates, in conjunction with the large active area of the detector means that it is well suited to collect high resolution synchrotron data.

Data collection and analysis (Imaging plate)

There are two main packages for analysing image plate data : MOSFLM (CCP4, unpublished program) and DENZO (Otwinowski, 1993). The latter was used for data processing of the GDH data. The flow chart depicts the general scheme of DENZO:

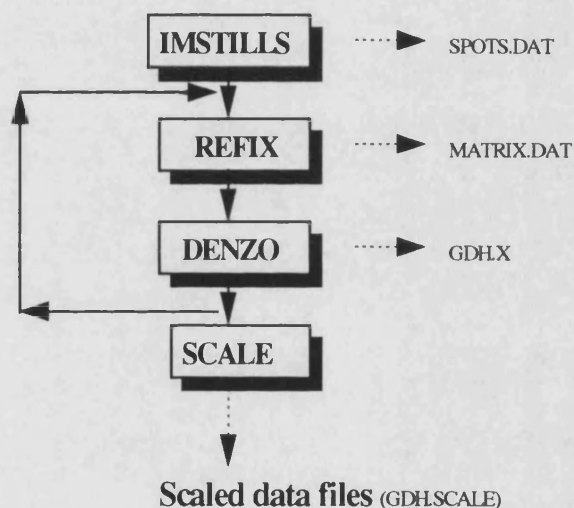


Fig. 3.2 Stages of the data reduction scheme of DENZO.

IMSTILLS: calculates the background of the image plate and finds the spots from up to 3 images. These images can be stills or oscillation images. **REFIX:** autoindexes the spots and determines the reduced cell based on the spots found from IMSTILLS. The predicted diffraction pattern (produced by OSCGEN of MOSFLM) can be compared with the experimental diffraction pattern to ensure that the correct orientation matrix

has been determined by REFIX. A standard deviation of spot position < 1 pixel, and that of the spindle position ($< 1/2$ the oscillation angle), indicates good data processing. **DENZO**: fits profiles to all the predicted spots on the images collected. The size of the box in which the spot and background are integrated is determined by the user. For every reflection a χ^2 test is calculated; values below 2 are indicative of good data processing. DENZO can refine all the instrumental and crystal parameters simultaneously, such as the radial and tangential offset of the detector, unit cell dimensions, the wavelength used, the crystal to detector distance and the direct beam co-ordinates. **SCALE**: merges all the processed images, applying the appropriate correction factors, and determines the mosaicity of the crystal. The mosaicity value, and the refined geometry parameters are put back into REFIX and DENZO to produce more accurately determined structure factors.

The unit cell

As mentioned in Chapter 2, the larger the crystal, the stronger the diffraction pattern. The macroseeding technique resulted in crystals of various sizes from which numerous native data sets of varying statistical quality were collected. Only results pertaining to the native data set used in subsequent analysis shall be discussed in detail. Individual data sets were merged using X-SCALE. The optimum data set was composed of 'in-house' data and synchrotron data collected from a number of crystals.

From the diffraction studies on smaller crystals, various crystallographic parameters were determined:

unit cell dimensions	a	b	c	α	β	γ
(Å)	81.85	120.08	88.38	90°	104.03°	90°

The crystal was classified in the monoclinic system, belonging to space group $P2_1$, as judged by systematically absent/weak axial reflections (the $0k0$ reflections are only present when $k=2n$).

The unit cell volume = 868647\AA^3 , molecular mass $\approx 40\text{Kd/monomer}$. If the asymmetric unit contains one tetramer, then V_m (volume occupied by 1 Dalton) = $2.71\text{\AA}^3/\text{Dalton}$ which is within the range given by Matthews for crystals. The solvent content of the crystal was estimated at 54% using the equation:

$$V_{\text{solv}} = 1 - \frac{1.23}{V_m}$$

In order to obtain a complete data set, at least 90° of data must be collected in the monoclinic system. However, the greater the redundancy in the data, the greater its statistical significance, hence efforts were made to collect as much native data as possible.

In-house data

Parameters: Generator settings: 35-40 kV, 75-80 mA

crystal detector distance = 12cm

exposure = 3-5minutes/frame

2θ swing ($\equiv\theta_c$) = $0-5^\circ$

The length of exposure and the θ_c swing was determined after observing the diffraction quality of the crystal.

A summary of the in-house data collection statistics is given in Table 3.1. This in-house data file is composed of 10 data sets, the majority of which display very good

correlation coefficients (>0.98 , highest = 0.994, lowest = 0.957). Despite the collection of a large amount of data (a seven-fold redundancy in the data) , the data were still not 100% complete at low resolution . This was probably due to the cusp of data that lies close to the rotation axis of the crystal; the crystal had one long dimension which tended to lie parallel to the capillary. The data does not extend greatly beyond 3.5Å and is essentially weak since there is a marked decrease in the completeness of the data as the signal/noise ratio is increased.

resolution (Å)	% completeness of data (I/σI)			No. of observed reflections	No. of unique reflections	No. of unique reflections ($>1\sigma$)
	1	2	3			
10.00	91.1	90.4	89.7	9301	866	862
6.00	91.8	89.1	85.9	35245	3093	3051
4.50	86.4	81.6	76.9	41829	5125	4976
4.00	82.2	75.2	68.2	21517	3663	3449
3.60	74.2	58.5	46.0	16800	4240	3884
3.5	68.4	47.1	30.5	3845	1354	1167

Table 3.1 Summary of data collection statistics from in-house data. Total number of unique

reflections = 18341, total number of reflections = 128537, overall $R_{\text{merge}} = 9.3\%$. See end of chapter

for glossary of terms used

Synchrotron data

The crystal to detector distance, and the direct beam co-ordinates were determined by collecting a paraffin wax image. The wavelength selected corresponded approximately to that of the bromine absorption edge, which was measured at the start of data collection.

A few problems were encountered in the data collection : arising mainly from the fact that the crystals were radiation sensitive, and the relatively low intensity of the beam

hitting the crystal (due to problems in steering the beam at the time of data collection). Thus although the crystals diffracted to 2.1Å, data were only processed to 2.9Å, and the data set presented (Table 3.2) was collected from 4 crystals. The synchrotron native crystals produced good REFIX solutions, but the crystals appeared not to be very isomorphous with the in-house crystals. A crystal previously exposed in-house gave a significantly different (i.e. a non-isomorphous cell when compared to the in-house data) unit cell using the synchrotron radiation. Therefore synchrotron data was processed whereby the unit cell was fixed to be isomorphous to that of the in-house data.

resolution (Å)	% completeness of data (I/σI)			No. of observed reflections	No. of unique reflections	No. of unique reflections (>1σ)
	1	2	3			
10.00	48.8	48.7	48.7	760	464	462
6.00	55.8	54.7	53.9	3209	1879	1856
4.00	58.0	56.0	53.8	10350	5928	5772
3.50	56.2	51.1	46.6	7176	4137	3901
3.20	52.5	43.7	36.6	6455	3820	3389
3.00	47.2	36.0	27.0	5569	3348	2772
2.90	43.7	30.6	22.0	3207	1975	1551

Table 3.2 Summary of the synchrotron data collection statistics. Total number of observed reflections = 36806, total number of unique reflections = 21551, overall $R_{\text{merge}} = 12.9\%$.

The mean correlation coefficients between the synchrotron data sets were > 0.95, which are reasonable values.

As a result of the time limitations at the synchrotron source, further data could not be collected, and this is reflected by the low completeness of data in the low resolution terms and the low redundancy of the data. Nevertheless, higher resolution data was collected (in comparison to the in-house data), which proved essential in map-interpretation (see chapter 6).

Final data file

The final data file was believed to represent the optimum native data file. It is a combination of the synchrotron data sets and some of the in-house data sets. As a result of poor correlation of the synchrotron data with some of the in-house data sets, the latter were discarded - some of the low resolution data was sacrificed to retain the higher resolution terms. The merging of these data files was a balance between trying to obtain a low R_{merge} , high correlation coefficients of the data sets, and higher resolution data.

The correlation coefficients between the in-house data sets and the synchrotron data sets varied considerably. The mean correlation coefficient was 0.93, however correlation coefficients as low as 0.88 were obtained.

As can be seen, the data were still relatively weak. The data were not very complete to 2.9Å at the 1σ cut-off; nevertheless the non-crystallographic symmetry introduced redundancy in the data, (see chapter 5), and interpretable maps were produced at 2.9Å (see chapter 6)

resolution (Å)	% completeness of data (I/I σ)			No. of observed reflections	No. of unique reflections	No. of unique reflections ($>1\sigma$)
	1	2	3			
10.00	86.4	85.8	85.4	6164	820	817
6.00	92.0	89.4	86.5	24584	3109	3059
4.50	91.6	87.3	82.6	31844	5406	5272
4.00	89.2	81.6	74.9	16899	3916	3744
3.50	84.6	70.6	60.3	24841	6373	5877
3.20	58.7	46.7	38.1	9684	4322	3793
3.00	47.2	36.0	27.0	5569	3348	2772
2.90	43.6	30.6	21.9	3207	1975	1549

Table 3.3 Summary of final data collection statistics. Total number of observed reflections =

122792, total number of unique reflections = 29269, overall $R_{\text{merge}} = 12.4\%$.

Non-crystallographic symmetry

Non-crystallographic symmetry (ncs) is symmetry within the asymmetric unit. GDH is a tetramer in solution, and appears to pack as one tetramer/asymmetric unit.

Consequently it is possible that GDH displays 222 ncs (three mutually perpendicular 2-fold axes) - a feature that is present in other dehydrogenases, such as glyceraldehyde 3-phosphate dehydrogenase (Buehner *et al.*, 1974) and some tetrameric alcohol dehydrogenases (Zhang *et al.*, 1993), or possibly the rarer 4 ncs symmetry as seen in the influenza virus neuraminidase (Varghese *et al.*, 1983).

Determination of the ncs is made via a search in Patterson space, whereby the Patterson function is rotated against itself to produce a self-rotation function. The rotation function was expressed in the spherical polar system, where ω and ϕ define the longitude and latitude respectively of the rotation axis and κ represents the rotation about this axis.

The relationship between the spherical polar axial system and the orthogonalisation system used is depicted in Fig. 3.3.

The POLARRFN program was used to compute the self-rotation function. The rotation function was sampled at 5° intervals on ϕ and ω , and 30° on κ . Peaks were found at section $\kappa = 180^\circ$ (in addition to the crystallographic 2-fold). A fine search about this κ axis ($176^\circ \Rightarrow 180^\circ$) was performed, with the most significant peaks being present at section $\kappa = 180^\circ$. No significant peaks were found on any other κ section, in particular the $\kappa = 90^\circ$, eliminating the 4-fold ncs.

A self rotation function of a monomeric protein, carbonic anhydrase, crystallising in the same space group $P2_1$ was computed as a control; no peaks were present at $\kappa = 180^\circ$, apart from the crystallographic 2-fold. A stereographic projection was used to display

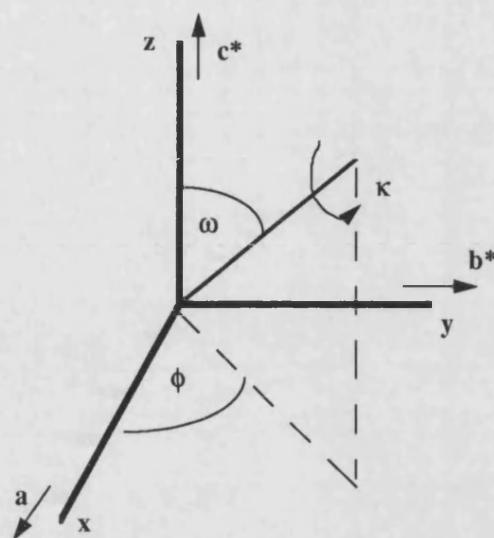


Fig. 3.3 The spherical polar system and the orthogonalisation matrix used.

the results (see fig 3.4). The direction cosines are output from the POLARRFN program. The angle between the peaks are calculated using the equation:

$x_a y_a z_a$ are the directional cosines of one peak and

$x_b y_b z_b$ are the directional cosines of another peak.

$$\cos \theta = x_a x_b + y_a y_b + z_a z_b$$

peak No.	ω (°)	ϕ (°)	% of origin peak	directional cosines		
				x_n $\sin\omega\cos\phi$	y_n $\sin\omega\sin\phi$	z_n $\cos\omega$
1	59.4	69.6	45	0.3007	0.8066	0.5089
2	147.6	44.0	33	0.3853	0.3715	-0.8447
3	78.9	332.3	33	0.8692	-0.4554	0.1925
4	59.4	290.4	45	0.3007	-0.8066	0.5089
5	147.6	316.0	33	0.3853	-0.3715	-0.8447
6	78.9	27.7	33	0.8692	0.4554	0.1925

Table 3.4 Vital statistics of the non-crystallographic peaks.

The angle between peak 1 & 2 = $90^\circ 49'$

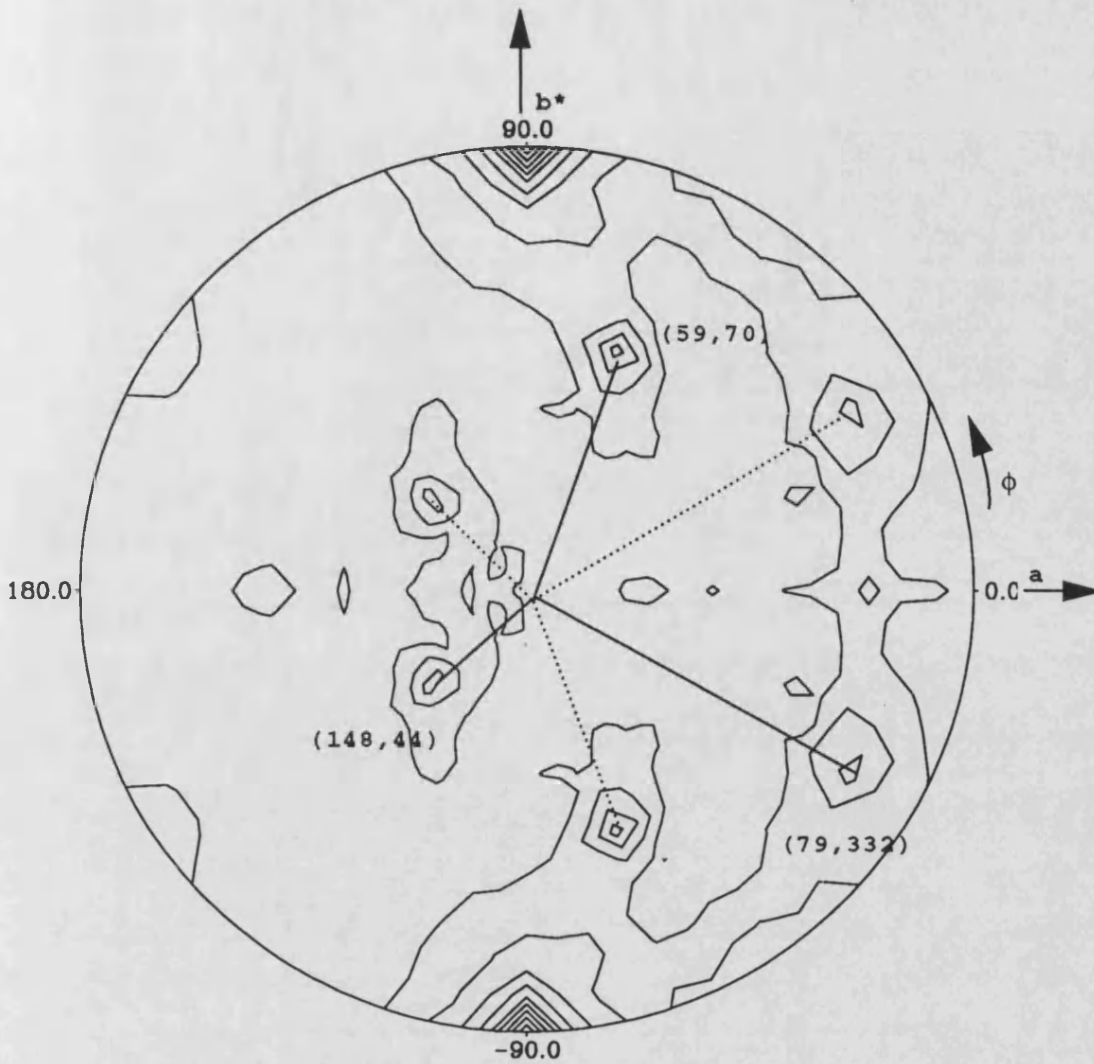
The angle between peak 1 & 3 = $90^\circ 27'$

The angle between peak 2 & 3 = $89^\circ 49'$

The angles between peaks 1→3 are approximately 90° , which is indicative of 222 ncs.

Peaks 4→6, which also display 222 ncs, are crystallographically related (180° rotation) to the 1→3 peaks.

The 222 ncs proved essential in solving the structure (see chapter 5)



SELF ROTATION OF GDH -native5 DATA
Section kappa = 180

Fig. 3.4 Stereographic projection of the self-rotation function. Integration radius of 30\AA , resolution range $10 \rightarrow 4.5\text{\AA}$, $\kappa = 180^\circ$. The three 222 ncs peaks joined at the origin by bold lines. The crystallographically related 222ncs is represented as dashed lines.

Glossary of terms used

$$R_{\text{merge}} = \frac{\sum_n |I_i - \langle I \rangle|}{\sum_n \langle I \rangle}$$

I_i = individual intensity measurement. $\langle I \rangle$ mean intensity measurement.

Chapter 4

ISOMORPHOUS REPLACEMENT

Introduction

In order to calculate an electron density map, both the amplitude and phase of each diffracted X-ray must be known. However, when collecting data only the intensities of the diffracted spots ($\text{amplitude}(F) \propto \sqrt{\text{intensity}(I)}$) are measured and all the phase information is lost - this is the phase problem. For small molecules, this can be overcome by direct methods. Theoretically, this can be applied to solving protein structures, however the complex nature of proteins and the lack of X-ray diffraction data to atomic resolution, in the majority of cases results in direct methods being an unlikely tool for phase determination. For protein crystallographers there are two classical methods for solving the phase problem

- **Molecular replacement** - which requires a model with a similar three-dimensional fold to that of the unknown to obtain a handle on the phases. Molecular replacement packages such as MERLOT (Fitzgerald, 1988), X-PLOR (Brünger, 1988) and AMORE (Nazava, 1992) are well documented as tools used in solving the structure of proteins.
- **Heavy atom isomorphous replacement** - requires the introduction of heavy atoms onto the protein. A substituted heavy atom will introduce new scattering

centres in the protein, and change the intensity of every reflection to some extent.

The position of these scattering centres can be found by the Patterson method.

From knowing the position of the heavy atoms and the amplitude of the native and heavy atom derivative structure factors, the phase can be estimated. Typically a combination of errors and phase ambiguity requires several different heavy atom derivatives being obtained.

There are no defined rules as to which heavy atom will make an effective derivative, however the major prerequisite of a good derivative is that the heavy atom binds specifically at a small number of sites and does not significantly perturb the structure of the protein. Finding a useful derivative involves a screening of heavy atom compounds; some reagents interact with specific amino acids (e.g. mercury with cysteine) whereas other reagents may interact electrostatically with protein side chains (e.g. the lanthanides).

There are two methods for introducing heavy atoms into a protein; firstly there is soaking heavy atoms into the crystal; secondly attempts can be made to co-crystallise proteins with the heavy atom.

Heavy atom derivatives

All derivative data were collected on the in-house area detector. Data processing, as previously described in Chapter 3 was employed, with the exception of a modified version of **XSCALE** (XX), which preserves the Friedel pair intensities. Subsequent data analysis was performed using the CCP4 suite of programs. The general scheme of derivative analysis is given in Fig. 4.1 and 4.2. The formatted reflection files (hkl

files) from XSCALE are converted into binary files (lcf) as used in the CCP4 suite

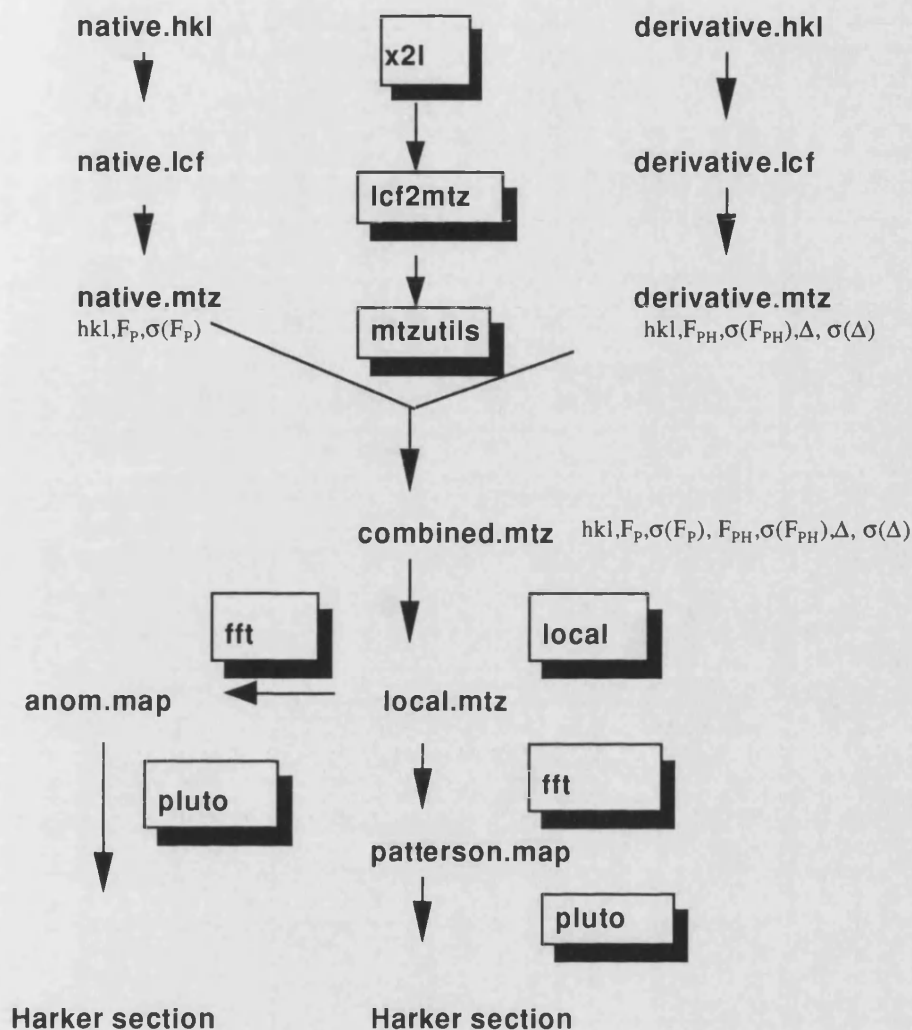


Fig. 4.1. A schematic representation of the stages involved in derivative data analysis. $\Delta = F_{PH(+)} - F_{PH(-)}$ (the anomalous difference).

(X2L). For the most recent CCP4 package, data reflection files are converted to MTZ binary format (LCF2MTZ) which can be read using MTZDUMP. This data file contains the reflection data records, where the data is stored as columns. Each column is given a column type and label. The first 3 columns are the Miller indices, while other column types such as, intensity (I), structure factor amplitude (F), anomalous

differences (D), standard deviation of a value (Q), a weight (W) and the phase probability coefficients can be added. **MTZUTILS**: combines the native and derivative data files. **LOCAL**: scales the native and derivative data sets together, and can scale the Friedel pair differences arising from the anomalous scattering by the heavy atom compounds. The program runs as two stages: first, the prescaling, where an overall scale factor is calculated using an anisotropic temperature factor. The second stage involves local scaling, where a reflection is scaled according to its environment. The fast Fourier transform (**FFT**): enables a rapid synthesis of an electron density map or Patterson map.

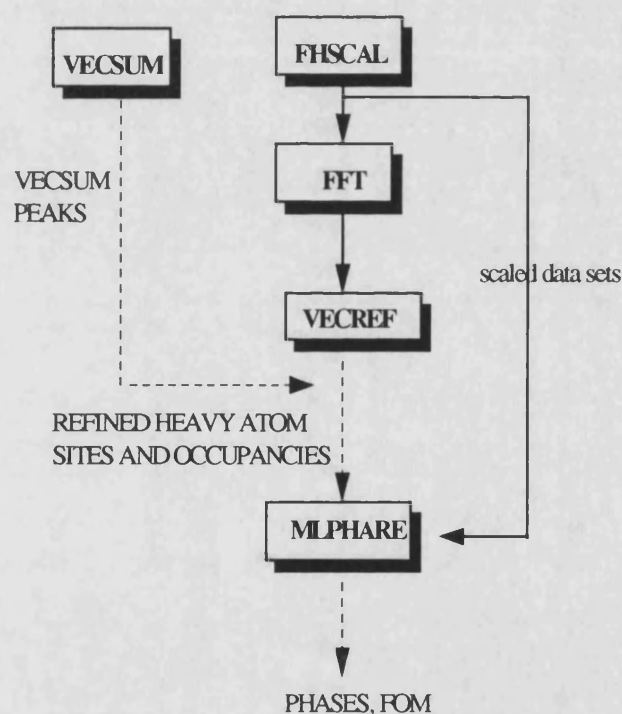


Fig. 4.2. Schematic representation of the stages involved in detection and refinement of the heavy atom sites.

In the context of fig. 4.1., it was used to calculate an isomorphous difference Patterson synthesis:

$$P(uvw) = \frac{2}{V} \sum_{hkl} (F_{PH} - F_P)^2 \cos 2\pi (hx + ky + lz)$$

The anomalous difference Patterson synthesis was the same as the above equation, except for the coefficients below were used:

$$(F_{PH(+)} - F_{PH(-)})^2$$

The sampling grid used for the FFT calculation (and plotting of the Patterson map) is dependent on the resolution of the data and the unit cell dimensions :-

$$grid \approx \frac{3}{\text{max. resol}^n} * (\text{unit cell dimensions})$$

PLUTO: plots the Patterson map in sections. The Harker section for space group $P2_1$ is $v=1/2$. **VECSUM**: automatically solves Patterson maps by assuming that every grid point in the asymmetric unit is a potential heavy atom site. Initially only the section $y=0$ need be searched because of the arbitrary origin along the y axis. Once the initial putative heavy atom site was found, the xyz co-ordinates were re-input into **VECSUM**, and further peaks sought. These peaks should correlate to the peaks on the Harker section.

The putative heavy atom sites are refined in **VECREP**. The refined heavy atom sites, and their respective occupancies are used to calculate the phases (**MLPHARE**). Prior to this refinement the native and derivative data sets are rescaled (**FHSCAL**) and a new correctly scaled isomorphous difference Patterson map is created (**FFT**) using the V-factor determined from **FHSCAL**. **MLPHARE** refines the heavy atom parameters, and uses this information to calculate phases (undergoing 10 cycles of refinement).

The phasing power (Ph_p) and the figure of merit (FOM) indicate how useful the derivative may be.

The Screen

As GDH packs as a tetramer in the asymmetric unit, it is reasonable to assume that a derivative would produce 4 (or multiples thereof) independent sites.

Numerous potential heavy atom compounds were screened in order to obtain derivatised GDH crystal(s). Table 4.1 represents a summary of the screen. Only results concerning the useful derivative obtained shall be described in detail.

The heavy atom compounds screened can be classified into 3 groups as judged by the lack of isomorphism and whether or not heavy atom sites were detected within the protein. Table 4.1 summarises all the heavy atom soaks attempted, and table 4.2 lists the unit cells relating to these soaks. Each class of derivative is discussed in turn.

compound	conc(mM)	soak time(hr)	isomorphous	derivative
PCMB	1	2	no	no
K_2PtCl_4	1	24	no	no
K_2PtCl_6	5	48	yes	very good
K_2PtCl_6	5	24	yes	good
K_2PtCl_6	5	72	no	no
$NaIr_3Cl_6$	5	72	yes	no sites
$Th(NO_3)_2$	5	72	yes	no sites
$NaAuCl_4$	5	96	yes	no sites
HgSCN	1	3	no	no
HgSCN+ glucose	0.1	1	no	no
IODINE/KI	0.5	1	no	no
UO_2SO_4	5	120	no	yellow crystal
UO_2SO_4	5	24	+/-	possible
$K_2Pt(CN)_4$	1	72	no	no

Table 4.1 Summary of the heavy atom screen.

unit cell (Å/°)	a	b	c	α	β	γ
native	81.85	120.08	88.38	90	104.03	90
PCMB	80.82	119.44	89.31	90	103.04	90
K ₂ PtCl ₄	81.15	117.07	86.66	90	103.53	90
K ₂ Pt(CN) ₄	81.81	120.76	89.89	90	103.76	90
NaIr ₃ Cl ₆	81.89	120.55	88.61	90	104.06	90
Th(NO ₃) ₂	81.54	119.85	88.21	90	103.94	90
NaAuCl ₄	81.97	120.08	88.60	90	104.10	90
UO ₂ SO ₄	81.98	118.35	87.65	90	103.68	90
K ₂ PtCl ₆	81.82	120.50	88.56	90	104.03	90

Table 4.2. The unit cells of some of the derivatised crystals.

Deleterious derivatives

These derivatives produced a non-isomorphous unit cell or disordered the crystal.

Regarding the derivatives, PCMB, K₂PtCl₄ K₂Pt(CN)₄, the unit cell in at least one dimension showed a change in cell dimensions greater than 1%.

Other derivatives, for example, HgSCN and Iodine completely disordered/cracked the crystal.

Potential derivatives

These derivatives were isomorphous with the native unit cell, however subsequent analysis showed that no heavy atoms were bound. Data collection statistics for each derivative is given in Table 4.3, and statistics relating to the potential usefulness of each derivative is given in tables 4.4 and 4.5. The isomorphous difference Patterson maps for each derivative is given in fig. 4.3 (a to d).

- **The iridium derivative**

The data collected is summarised in table 4.3. The change in cell dimensions were

(determined from table 4.2) : % Δ in a = 0.04%, % Δ in b = 0.40%, % Δ in c = 0.26%

and $\% \Delta$ in $\beta = 0.03\%$.

Information relating to the derivative analysis is given in table 4.4 and 4.5. There were no significant peaks 2 peaks at 3.6σ) found on the Harker section.

Derivative	resolution (Å)	completeness of data (%)	No. of observed reflections	No. of unique reflections
Iridium	10	55.1	1086	527
	6	57.8	3569	2040
	4.5	51.3	3706	3238
	3.6	18.6	2183	2097
Thorium	10	84.7	1954	807
	6	85.7	7415	2995
	4.5	74.7	7113	4694
	3.6	17.0	941	834
Gold	10	58.1	1075	559
	6	62.4	4538	2309
	4.5	32.6	4228	2349
Uranyl	10	70.4	1639	677
	6	75.6	6747	2675
	5.0	67.3	5316	2304

Table 4.3 Summary of the data collection statistics. Total R-merge = 4.5%, 4.5%, 9.3% and 8.1%

for the Iridium, Thorium, Gold and Uranyl derivatives respectively.

These two putative sites were put into MLPHARE to calculate phases, and produced a very low FOM (0.16), and a low phasing power (0.5). A cross phase difference Fourier (using the phases derived from the platinum derivative) was also calculated. The difference map was searched for peaks ($>2\sigma$). No peaks were found that correlated to the Iridium Harker section. At this stage in the analysis, it was decided that the potentiality of the iridium derivative was not worth pursuing.

	Fractional intensity difference	residual.	max. peak value (σ)	No. of putative peaks
Iridium	24.1	8.4	3.6	2
Thorium	21.0	7.0	3.7	4
Gold	26.7	8.6	3.3	4
Uranyl	72.0	26.0	5.0	4

Table 4.4. Summary of the scaling statistics and the number and significance of putative heavy

atom sites found using VECSUM. For glossary of terms see end of Chapter.

- **thorium derivative**

The data collected is summarised in table 4.3. The change in cell dimensions were

(determined from table 4.2): $\% \Delta$ in $a = 0.38\%$, $\% \Delta$ in $b = 0.19\%$, $\% \Delta$ in $c = 0.18\%$

and $\% \Delta$ in $\beta = 0.08\%$

4 peaks were found (using VECSUM) on the Harker section (maximum peak value of

3.7σ), whose occupancies refined to zero in VECREF. Hence thorium was not a

derivative.

Derivative site	fractional co-ordinates	relative occupancy	residual	correlation
Ir1	0.10, 0.0, 0.16	1.17	0.70	0.77
Ir2	0.30, 0.09, 0.45	1.07	0.71	0.75
Au1	0.08, 0.0, 0.11	6.01	0.64	0.88
Au2	0.90, 0.20, 0.89	3.98	0.64	0.86
Ur1	0.0, 0.0, 0.11	1.5	0.67	0.89
Ur2	0.08, 0.06, 0.75	1.7	0.69	0.79
Ur3	0.0, 0.50, 0.89	1.5	0.67	0.89
Ur4	0.93, 0.08, 0.50	1.8	0.66	0.83

Table 4.5 Represents further analysis of the putative sites from table 4.4, *e.g.* Ir1 and Ir2 represent

the 2 sites found from VECSUM. A high correlation and low residual is indicative of a good

derivative. A low residual and high correlation is indicative of a good derivative.

- **gold derivative**

The data collected is summarised in table 4.3. The change in cell dimensions were (determined from table 4.2): $\% \Delta$ in $a = 0.14\%$, $\% \Delta$ in $b = 0.01\%$, $\% \Delta$ in $c = 0.26\%$ and $\% \Delta$ in $\beta = 0.02\%$.

Using VECSUM, 4 possible peaks were found, 2 of which when refined in VECREF, gave very low occupancies and high residuals. The two remaining peaks displayed reasonable statistics (table 4.4 and 4.5), but produced a poor phasing power and a low FOM when input into MLPHARE. In addition a cross phase difference Fourier (using the platinum phases) gave no peaks that could be correlated to the gold derivative Harker section. At this stage in the analysis, it was decided that the potentiality of the gold derivative was not worth pursuing

- **uranyl derivative**

GDH crystals soaked in 5mM uranyl sulphate for 5 days turned a yellow colour, however these crystals were disordered. This evidence of uranyl binding indicated its potential as a derivative. A shorter soak was undertaken (24hr), however the cell was non-isomorphous with respect to the native (table 4.2). Nevertheless, the data were analysed, and a Patterson map calculated, in which significant peaks were present ($>5\sigma$). Some of these peaks were on the edges of the Patterson map - which is indicative of a non-isomorphous derivative. The change in cell dimensions were (determined from table 4.2): $\% \Delta$ in $a = 0.13\%$, $\% \Delta$ in $b = 1.44\%$, $\% \Delta$ in $c = 0.83\%$ and $\% \Delta$ in $\beta = 0.33\%$. The large fractional intensity difference and residual was indicative of a non-isomorphous derivative.

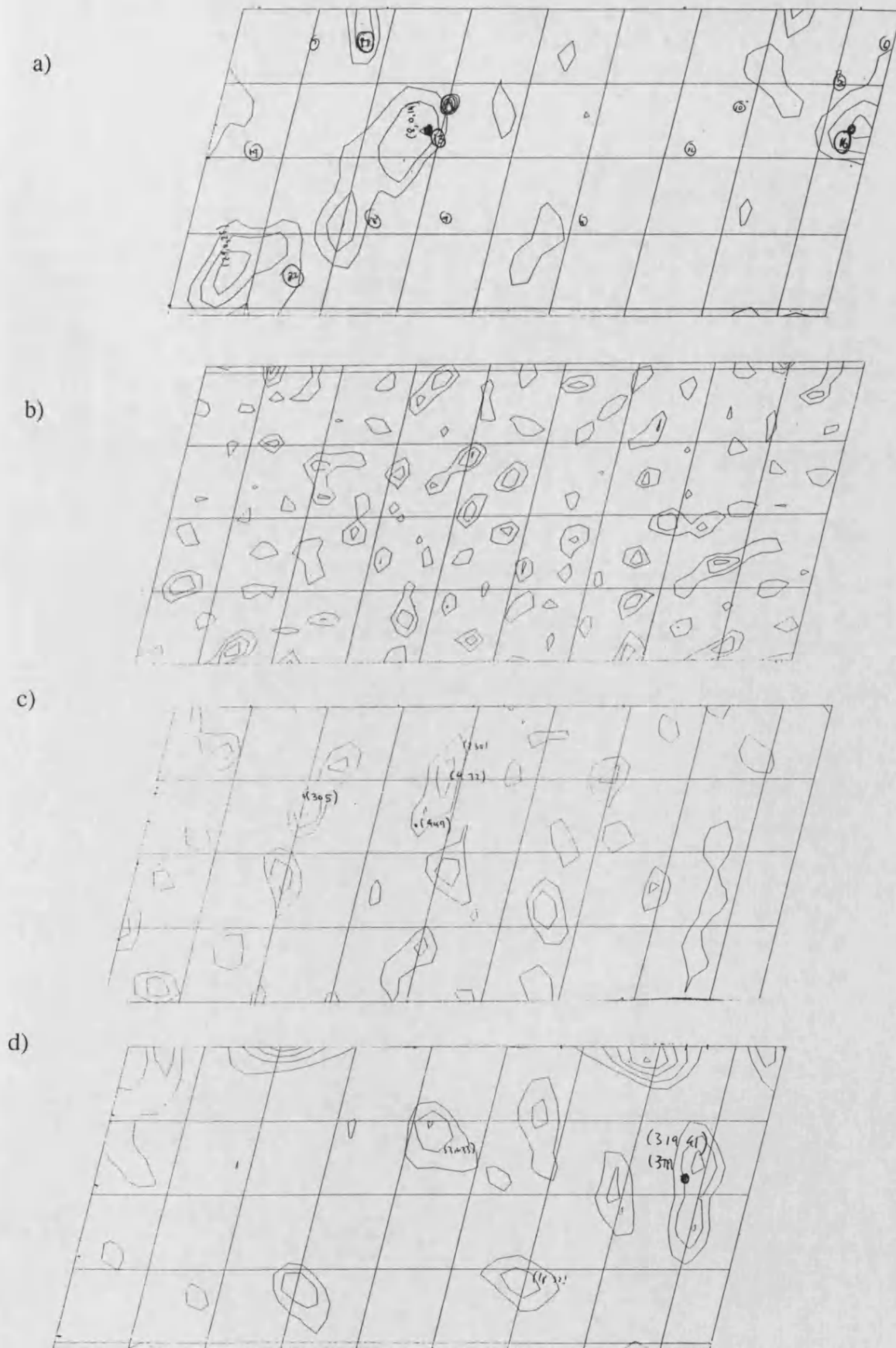


Fig. 4.3 Isomorphous difference Patterson maps at the Harker section ($v=1/2$) for a) Iridium b) Thorium c) Gold d) Uranium.

4 peaks were found using VECSUM, which were refined in VECREF (table 4.5)

These uranyl sites were input into MLPHARE, producing a FOM = 0.33, phasing power = 1.1 and a Cullis R = 0.19 (all data to 6Å). Therefore there were indications that the uranyl compound may prove to be a useful derivative. To ascertain further whether there was any ordered uranium atoms in the crystals, two cross phase difference Fourier's were calculated with coefficients:

$$1) FOM(F_{US} - F_{NAT})e^{i\alpha} (SIR_{Pt} \text{ phases})$$

$$2) FOM(F_{PT} - F_{NAT})e^{i\alpha} (SIR_{US} \text{ phases})$$

Both difference maps were searched for peaks $>3\sigma$ (using PEAKMAX). Peaks corresponding to 3 of the 4 uranyl sites were found using difference map 1, furthermore 4 peaks corresponding to the platinum derivative were found using difference map 2.

Obviously there is some element of truth in this uranyl derivative, however due to the large degree of non-isomorphism, and the concomitant success of the ncs averaging, it was decided not to include the uranyl data in further analysis.

The successful derivative (the platinum derivative)

The GDH crystals appeared to be sensitive to the duration of the platinum soak. The optimum soak conditions were determined to be 48hr (5mM K_2PtCl_6), whereas a 24hr soak still produced a useful derivative, but the occupancies of the platinum sites were halved. The change in cell dimensions were (determined from table 4.2): $\% \Delta$ in a = 0.04%, $\% \Delta$ in b = 0.35%, $\% \Delta$ in c = 0.20% and $\% \Delta$ in β = 0.0%. In contrast, a 72hr platinum soak produced a non-isomorphous unit cell. Data were collected from 4

different platinum crystals (designated PT1, PT2, PT3 and PT4). PT1, PT3 and PT4 represent the 48 hr platinum soak; PT2 represents the 24hr platinum soak. As it is unwise to merge different derivative data sets from separate crystals, each platinum soaked crystal was analysed as a separate derivative. PT1 was the first data set from which the Patterson maps were created, and heavy atom sites determined, refined and initial phases calculated. The subsequent platinum data sets, after initial inspection of the Patterson maps, were used in MLPHARE. Results relating to the detection of the heavy atom positions, and refinement of these positions shall only be discussed for PT1.

Data collection

Below is a tabulated summary of the 4 platinum data sets collected

Data set	resolution (Å)	completeness of data (%)			No. observed reflections	No. of unique reflections
		1 σ	2 σ	3 σ		
PT1	10	80.9	87.4	86.4	3332	847
	6	90.7	84.5	78.2	11958	3148
	4.4	73.8	62.4	52.3	12540	5283
	4.2	55.4	40.7	30.0	1552	1025
	4.0	43.7	28.9	19.3	1474	1052
PT2	10	67.4	66.7	66.3	1282	650
	6	71.8	67.4	63.2	4229	2497
	4.5	64.1	55.4	49.0	4471	3982
	4.0	47.6	36.8	29.9	2403	2267
PT3	10	55.7	55.2	53.7	1281	531
	6	58.6	50.3	44.2	5200	2117
	4.5	39.2	24.5	16.1	5082	2938
PT4	10	55.6	54.4	53.6	1206	532
	6	60.3	53.4	46.8	5030	2164
	4.5	59.2	45.2	33.8	8724	3959
	4.0	36.4	21.6	13.6	3824	1973

Table 4.6 Summary of data collection statistics of the 4 platinum data sets. $R_{\text{merge}} = 9.8\%$, 4.0% ,

8.6% and 8.3% for PT1, PT2, PT3 and PT4 respectively.

Derivative analysis

For statistics of the local scaling see fig. 4.4. The 4.5Å isomorphous difference Patterson map contained 4 significant peaks that were greater than 5σ (fig. 4.5). The positions of these peaks were determined using VECSUM, which were subsequently correlated to the Harker section peaks.

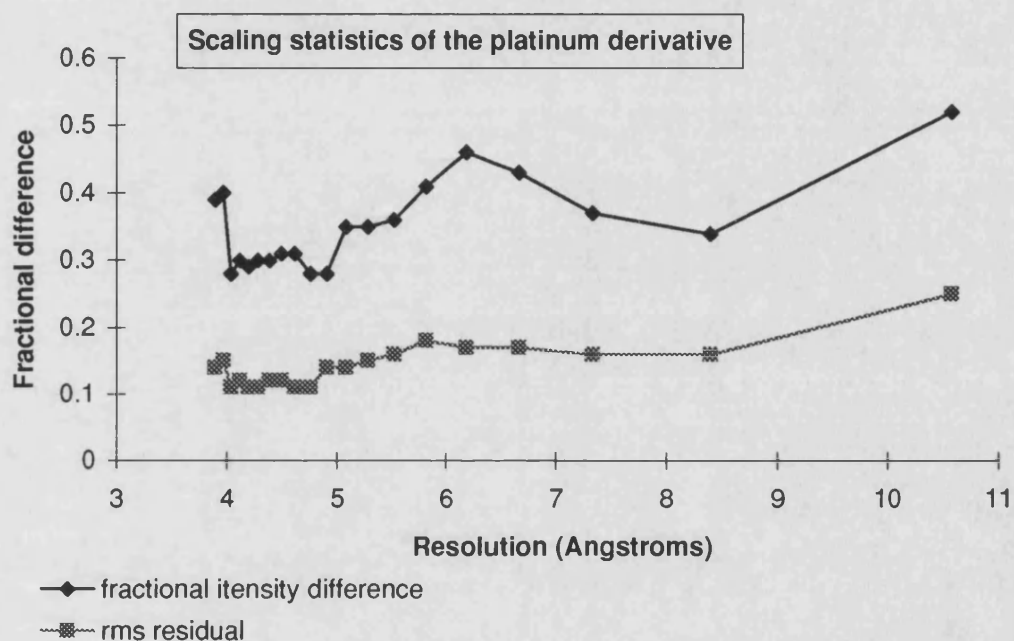


Fig. 4.4. A graph displaying the fractional intensity difference and the rms residual as a function of resolution. The platinum/native data scaling appears to be independent of resolution.

Vecsum grid co-ordinates			Harker peaks		fractional co-ordinates		
u	v	w	u=2x	w=2z	x	y	z
8	0	19	16	38	0.15	0	0.32
18	6	3	36	6	0.33	0.08	0.05
1	22	33	2	66	0.02	0.28	0.55
21	12	25	42	50	0.39	0.15	0.42

Table 4.7 The 4 heavy atom sites found and the fractional co-ordinates of these sites. The GRID:

54 80 60.

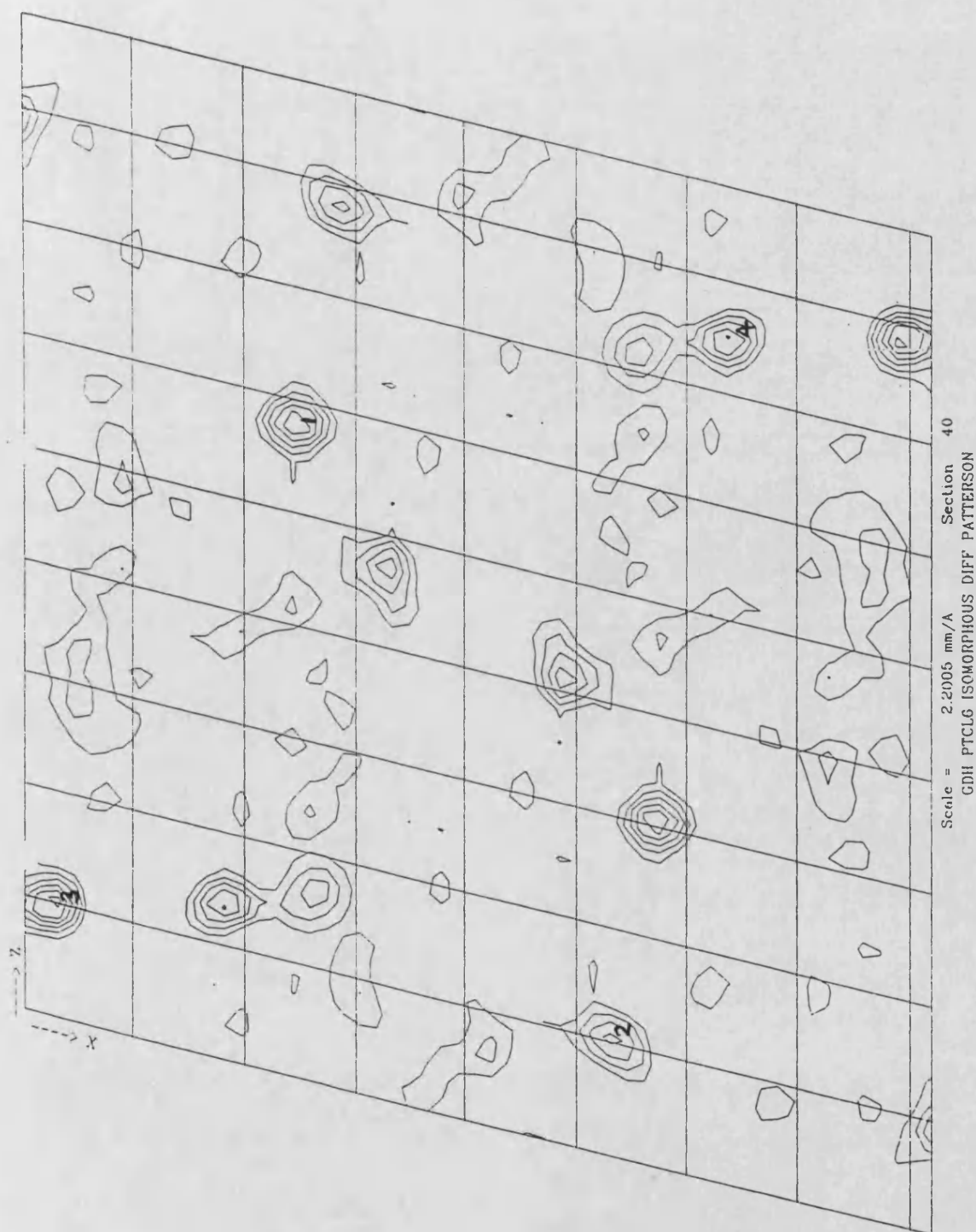


Fig. 4.5 4.5Å Isomorphous difference Patterson map at the Harker section ($v=1/2$) for the Platinum derivative. GRID: 54 80 60. The 4 peaks are indicated.

These putative platinum sites were refined in VECREF (table 4.8)- the refined fractional co-ordinates did not deviate greatly from the co-ordinates estimated from the Patterson map. All four sites exhibited similar occupancies, reasonably low residuals, and correlated very well. The platinum sites were also refined with the additional scattering factors attributed to the chlorine atoms (PtCl_4^{2-}), which gave similar statistics apart from the occupancies, which had halved (1.26, 1.23, 1.18, 1.14 for the 4 sites respectively), as would be expected.

Using these 4 main platinum peaks, the platinum cross vectors were calculated. Cross vector peaks on the Patterson sections 6, 22, 16, 13 were found. The calculated cross vector peaks correlated with the peaks found on the Patterson sections, which validated the existence of the platinum sites.

sites	x	y	z	relative occupancy	residual	correlation
1	0.1482	0.0019	0.3287	2.25	0.461	0.894
2	0.3271	0.0799	0.0564	2.21	0.470	0.880
3	0.0138	0.2795	0.5565	2.09	0.485	0.877
4	0.3905	0.1562	0.4232	2.04	0.521	0.845

Table 4.8 Heavy atom site refinement parameters. A low residual and high correlation is indicative of a good derivative.

Anomalous patterson map

There was no significant anomalous signal found on this Patterson map, furthermore the peaks on the anomalous Patterson did not correspond to the isomorphous difference Patterson peaks. These observations imply that there were no useful anomalous data.

Phase calculation

The phases were calculated using MLPHARE. Initially, phases were calculated using the PT1 data set. Subsequently, more platinum derivatives were obtained (PT2, PT3 and PT4), and better quality native data sets were collected. Table 4.9 represents phasing statistics regarding PT1 (to 4.5Å), whereas table 4.10 represents information from the 4 Pt crystals and the optimum native data set (to 4Å). The overall FOM for PT1 was 0.342 and 0.298 to 4.5Å and 4.0Å respectively (for acentric reflections).

resolution	N-ref	DISO	LOC	Php	Cullis R
14.14	189	445	317	1.2	0.24
10.54	335	338	195	1.8	0.22
8.77	565	280	149	2.1	0.22
7.25	809	230	118	2.2	0.16
6.32	1128	202	117	1.8	0.16
5.59	1461	175	121	1.4	0.17
5.00	1772	172	140	1.0	0.19
4.50	1739	177	160	0.6	0.20
total	7998	213	147	1.3	0.19

Table 4.9 Phasing statistics of the PT1 derivative. For explanation of terms see Glossary at end of Chapter. Diso, isomorphous difference; LOC, lack of closure; Php, phasing power; N-ref, number of reflections. Data to 4.5Å resolution.

	PT1	PT2	PT3	PT4
N ^o reflections	8995	7732	4882	7685
DISO	212	129	313	222
LOC	132	97	234	163
Cullis R	0.17	0.17	0.19	0.15
Php	1.6	1.1	1.1	1.2

Table 4.10 Phasing statistics of the 4 independent derivatives derived from MLPHARE. Data to 4Å resolution (PT3 to 4.5Å resolution).

PT1 appears to be the most useful data set in terms of the phasing power, and from tables 4.9, 4.10 & 4.11, the advantage of the other derivative data sets in improving

Resolution (Å)	13.33	10.00	8.00	6.67	5.71	5.00	4.44	4.0
FOM	0.655	0.705	0.691	0.668	0.603	0.520	0.362	0.274

Table 4.11 Figure of merit (FOM) (acentric reflections) as a function of resolution for the 4 Pt

crystals. Overall FOM = 0.46.

the overall FOM for the acentric reflections, and the phasing efficacy of PT1 can be seen. The FOM and the phasing power decreased as the resolution increased, which again underlines the weakness of the data collected. The overall FOM for the centric reflections were much higher than the acentric reflections (0.71 in comparison to 0.46 respectively), however as very few centric reflections were collected, these results are not shown.

Minor sites

Using the platinum phases the difference Fourier was calculated:

$$FOM (F_{Pt} - F_{nat}) e^{i\alpha (SIR_{Pt} \text{ phases})}$$

No significant peaks were found on the difference map (PEAKMAX), indicating that there were no minor platinum sites present.

Glossary of terms used

$$R_{\text{merge}} = \frac{\sum_n |I_i - \langle I \rangle|}{\sum_n \langle I \rangle}$$

I_i = individual intensity measurement. $\langle I \rangle$ mean intensity measurement.

$$\text{Phasing power} = Rms \left(\frac{f_H}{\text{lack of closure}} \right) \quad \text{Cullis R} = \frac{\text{lack of closure}}{\text{isomorphous difference}}$$

$$\text{Lack of closure} = \sum \left| |F_{PH} \pm F_p| - f_H \right| \quad \text{Isomorphous difference} = \sum \frac{|I_{nat} - I_{deriv}|}{\sum \langle I \rangle}$$

$$Rms \text{ Lack of closure} = \left[\frac{\left(\sum \left| |F_{PH} \pm F_p| - f_H \right|^2 \right)}{n} \right]^{0.5} \quad Rms f_H = \left(\frac{\sum f_H^2}{n} \right)^{0.5}$$

I = diffraction intensity (I_{nat} , I_{deriv} intensity of native and derivative diffraction

respectively); f_H = calculated heavy atom structure factor. F_p = structure factor

amplitude of native crystals. F_{PH} = structure factor amplitude of derivative crystals.

Sums are over all reflections n.

Chapter 5

RAVING

Real space introduction

As a result of the phase ambiguity, electron density maps from a single derivative (SIR maps) are generally uninterpretable as they contain the true electron density map superposed on a map of random noise. Generally, to solve the phase problem, and obtain an interpretable map at least two derivatives are needed (or good anomalous data from one derivative). Nevertheless, information regarding the protein boundary within the unit cell may be obtained from an SIR map. In addition, there are density modification procedures, such as Wang solvent flattening (Wang,1985) and SQUASH (Zhang, 1993) that attempt to improve the quality of the electron density maps.

At this juncture, given that phases were to 4\AA , only solvent flattening was applied, whereby the electron density within the molecular boundary is enhanced relative to the solvent region and any negative electron density is removed. Solvent flattening has been shown to be a very effective aid in delineating the molecular boundary (Wang, 1985).

Approximate boundary delineation

The SIR map indicated where the GDH tetramer may reside within the unit cell.

The solvent content of the GDH crystal is approximately 54%, however in the solvent flattening procedure, a solvent content of 49% was chosen to reduce the truncation of any surface loops. After 8 cycles of Wang solvent flattening, the boundaries of the tetramer were enhanced. This solvent flattened map, in conjunction with a skeletonised map (using BONES (Greer, 1974; Jones *et al.*, 1991)) enabled the approximate boundaries of the tetramer to be determined within the asymmetric unit. From a visual inspection of the map and bones, the fractional xyz limits of a box surrounding the complete tetramer were estimated to be (fig. 5.1):

$$\begin{array}{ccc} \text{x} & \text{y} & \text{z} \\ -0.33 \text{ to } 0.75 & 0.30 \text{ to } 0.80 & 0.25 \text{ to } 1.35 \end{array}$$

Non-crystallographic symmetry operators

From the self-rotation function, tetrameric GDH was shown to display 222 ncs; from the Patterson synthesis, 4 platinum sites were found per asymmetric unit.

Consequently it would be reasonable to assume that there were 4 platinum sites/tetramer, one platinum site/monomer and furthermore, each platinum would be substituted at the same region on each monomer. Hence the platinum sites additionally should display 222ncs (see fig. 5.2)

Firstly, the co-ordinates of the platinum sites related to a single tetramer within the defined boundary (above) had to be determined. The fractional co-ordinates of the 4 platinum sites were transformed into Cartesian co-ordinates (using the Brookhaven orthogonalisation convention). From a visual inspection (in FRODO) of the four platinum sites and their symmetry related counterparts, the 4 sites within the defined protein boundary in the solvent flattened SIR map were established. Their co-ordinates are listed in table 5.1.

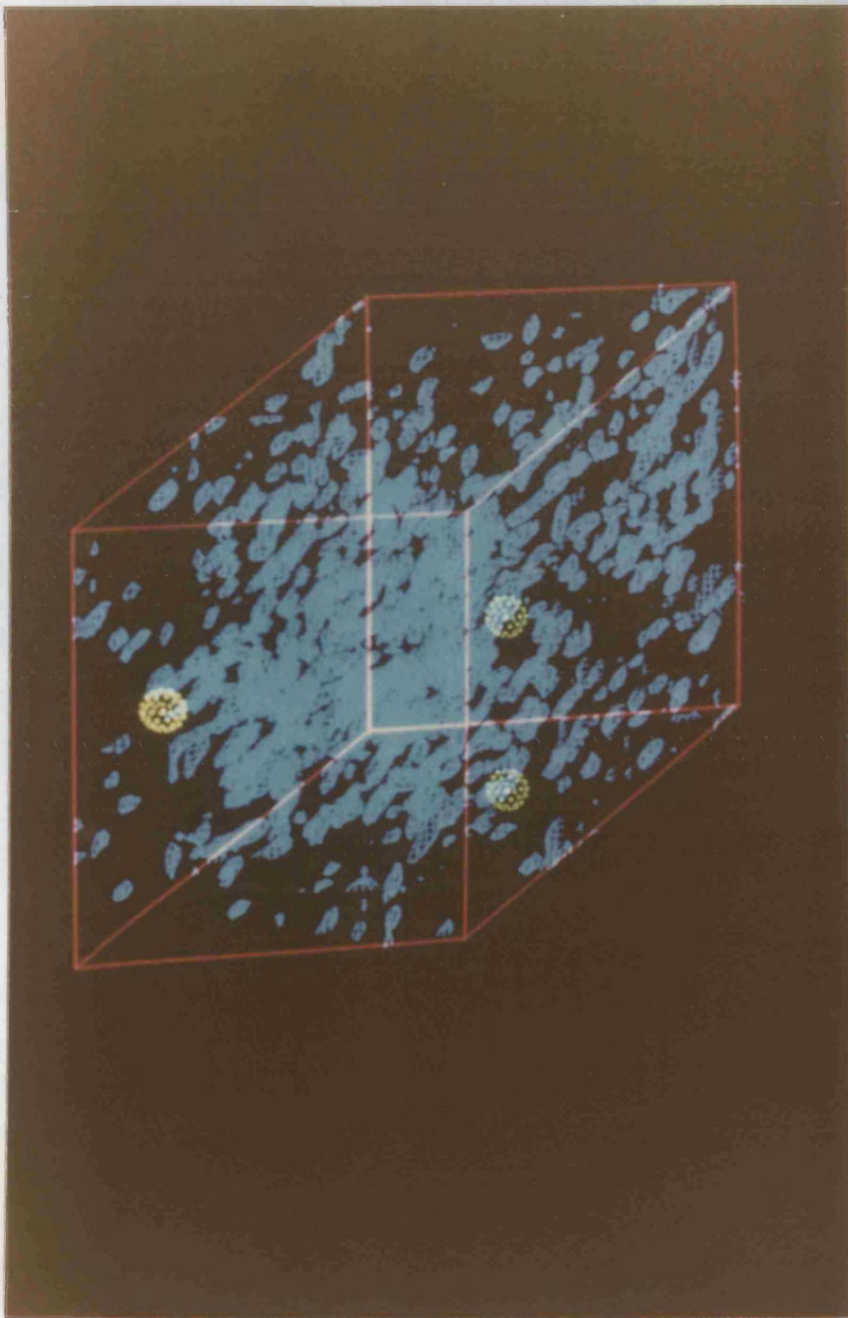


Fig. 5.1. Approximate tetramer boundary within the unit cell. The green spheres represent the platinum sites.

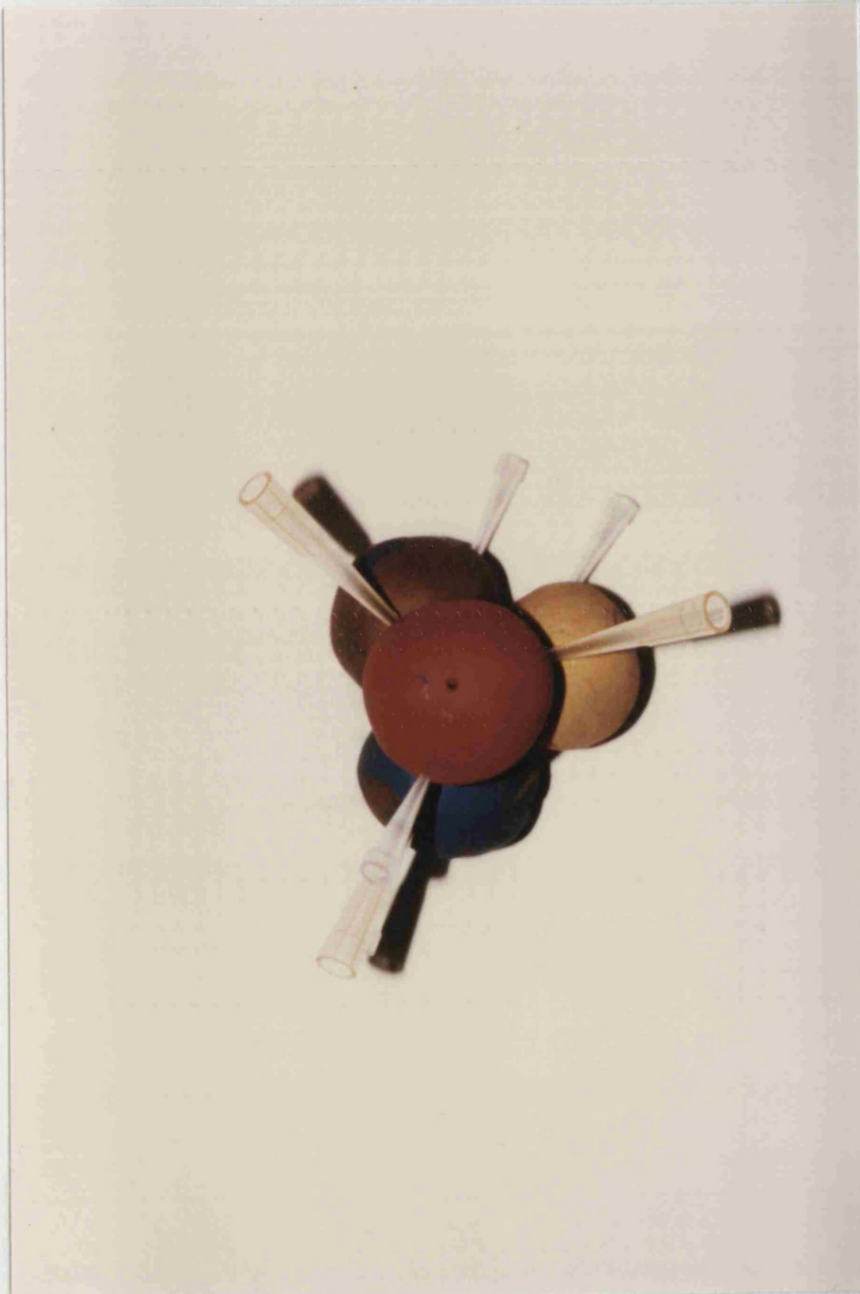


Fig. 5.2. Tetrameric model of GDH displaying the non-crystallographic symmetry.


platinum sites	co-ordinates (Å)		
	x	y	z
1	-26.49	60.40	57.53
2	34.88	69.65	80.86
3	-10.77	33.62	47.67
4	37.65	78.77	49.47

Table 5.1 The co-ordinates of the 4 platinum atoms associated with one tetramer.

From these co-ordinates the centre of mass (the origin) of the tetramer was calculated to be : **8.82, 60.61, 58.88 (Å)**. From the direction cosines of the 222 ncs and the centre of mass of the tetramer, the rotation and translation matrices relating one GDH monomer to any other monomer were derived (see fig. 5.3).

The operation below equates point x_b (on monomer b) to point x_a (on monomer a):

$$x_b = R_b x_a + T_b$$

where R_b is the rotation element and T_b is the translation element. The rotation component was derived from the ncs operators and the translational component from the ncs operators and the centre of mass of the tetramer. The calculations were done using a  *Garry's jiffy*. The 3 ncs operators are listed below.

monomer A to B rotation matrix:

$$\left\{ \begin{array}{ccc} -0.820 & -0.484 & 0.306 \\ -0.484 & 0.302 & -0.821 \\ 0.306 & -0.821 & -0.482 \end{array} \right\} \text{ translation vector: } \{27.40 \quad 94.96 \quad 134.34\}$$

monomer A to C rotation matrix:

$$\left\{ \begin{array}{ccc} -0.704 & -0.286 & -0.650 \\ -0.286 & -0.724 & -0.628 \\ -0.650 & 0.628 & 0.427 \end{array} \right\} \text{ translation vector: } \{70.66 \quad 70.01 \quad 1.39\}$$

monomer A to D rotation matrix:

$$\begin{Bmatrix} 0.510 & 0.793 & 0.335 \\ 0.793 & -0.584 & 0.176 \\ 0.335 & 0.176 & -0.930 \end{Bmatrix} \text{ translation vector: } \{-63.42 \quad 78.67 \quad 99.81\}$$

The validity of these ncs operators were checked using the platinum co-ordinates.

Each ncs operator produced a non-crystallographically-related platinum site.

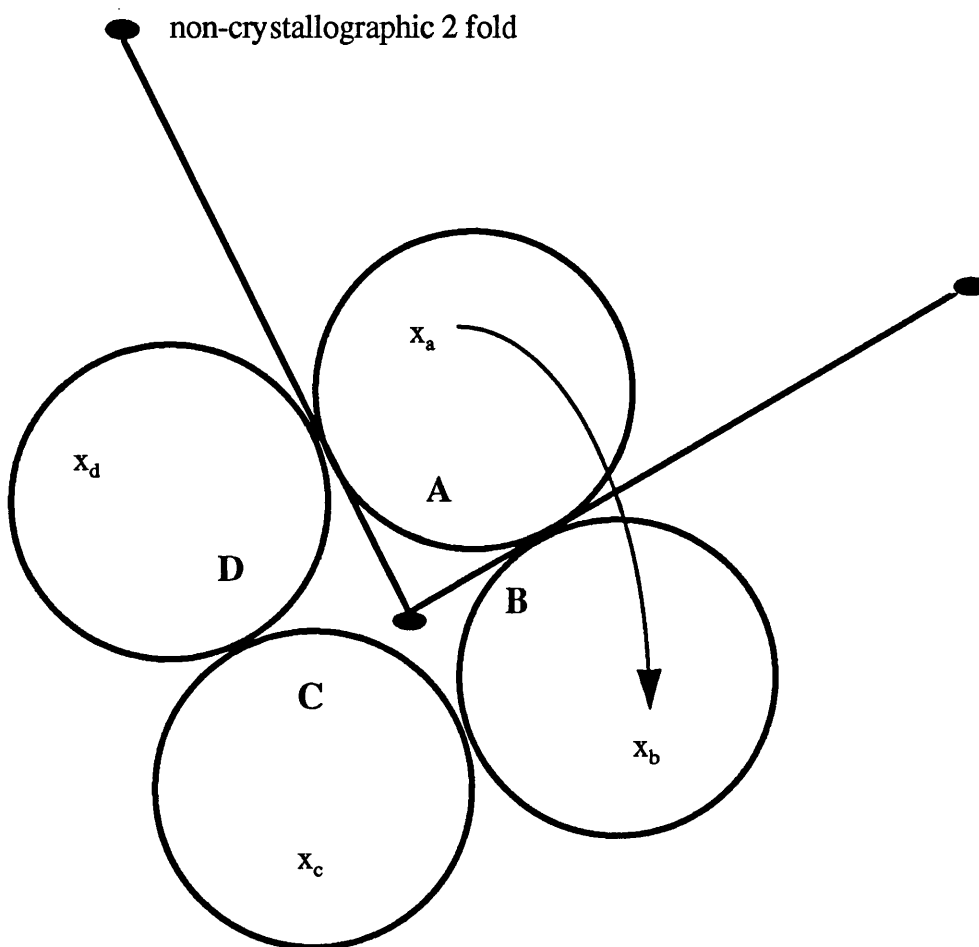


Fig. 5.3 Schematic representation of the 4 GDH monomers(A,B,C,D) and the rotation required to equate point x_a onto x_b .

Symmetry averaging introduction

Symmetry within the asymmetric unit is termed non-crystallographic symmetry (ncs).

There are two types of ncs, proper and improper. The former arises when a protein

displays point group symmetry, for example some tetrameric dehydrogenases display 222 point symmetry (Zhang *et al.*, 1993). Improper ncs is where the molecules in the asymmetric unit do not exhibit point symmetry. Here the directionality of the transformation is of paramount importance in defining the ncs operator. Whatever the nature of the ncs, redundancy is introduced into the data which can be exploited, by using the molecular averaging techniques, to increase the signal and reduce the errors in the initial unaveraged map.

The theoretical potential of ncs averaging was first formulated in reciprocal space (Rossmann & Blow, 1962), however only real space ncs averaging appears to have been implemented practically (Bricogne, 1976).

Symmetry averaging has proved to be a powerful technique in improving the quality of electron density maps, and indeed solving the structures of a number of proteins (Jones, *et al.*, 1991; Rossmann, 1990), including that of aldose reductase, from which only a single derivative was obtained (Tête-Favier *et al.*, 1993).

Glucose dehydrogenase packs as a tetramer in the asymmetric unit, with the tetramer displaying strict 222 ncs. When proteins displaying proper ncs are averaged, it is not necessary to define the boundaries of the non-crystallographic asymmetric unit (the monomer). For GDH only the boundaries of the tetramer need to be delineated; the 222 ncs will apply everywhere within the molecular envelope containing the 4 monomers. The electron density at every point within the envelope (mask) can be averaged among all 4 related points. In improper ncs, the boundaries of the non-crystallographic asymmetric unit are required.

There are three fundamentally important parameters required for symmetry averaging

- A set of phases from which an electron density map can be calculated. The quality of the initial data set, and in particular the completeness of data in the **low resolution terms** have been shown to be vital (Tête-Favier *et al.*, 1993). The low resolution terms determine the general positioning of the secondary structural elements of a protein. The quality of the initial phases are also important, however successful symmetry averaging from an uninterpretable SIR map has been achieved (The GDH structure !, John *et al.*, 1994).
- A molecular envelope (**mask**) in which the averaging is to be employed. The choice of mask is of vital importance for symmetry averaging. Too small a mask will result in truncation of electron density, probably at the surface loops; conversely too large a mask will cause clashes with crystallographically related molecules (Kleywegt & Jones, 1994).
- A description of the **ncs operators**. The greater the extent of the ncs, the more powerful the averaging technique. The ncs operators determine which points are equivalent within the mask, and consequently which points to average. An accurate determination of the ncs operators is therefore a prerequisite.

The RAVE scheme

The programs employed in the real space averaging were from the Uppsala group (Kleywegt and Jones, 1994): **RAVE** and **MAMA**. One cycle of averaging needs to be completed, to produce an initial averaged map, prior to 1 round of symmetry averaging (fig. 5.4). The calculated structure factors and calculated phases (F_{calc} , PHI_{calc}) derived from the initial averaged map are obtained using SFALL, and combined with the observed structure factors (F_{obs} , $\sigma(F_{\text{obs}})$). The observed and calculated structure factors

are scaled using a least squares scaling procedure (RSTATS), and FFT calculates a $2F_0$ - F_c electron density map. The map is then averaged (AVE) within the defined mask using the appropriate ncs operators, and subsequently expanded (EXPAND) over the whole unit cell. The expanded/averaged map is then reinput into SFALL for another cycle of symmetry averaging. Five cycles of averaging constituted one round of symmetry averaging.

Although the reflection data need not be expanded to the space group P1 prior to averaging, in the case of GDH the data were expanded to P1 and the subsequent averaging was in P1. The reason being that the initial phases were not good enough to adequately define the molecular boundary, and so erring on the side of caution the data were expanded to the whole unit cell.

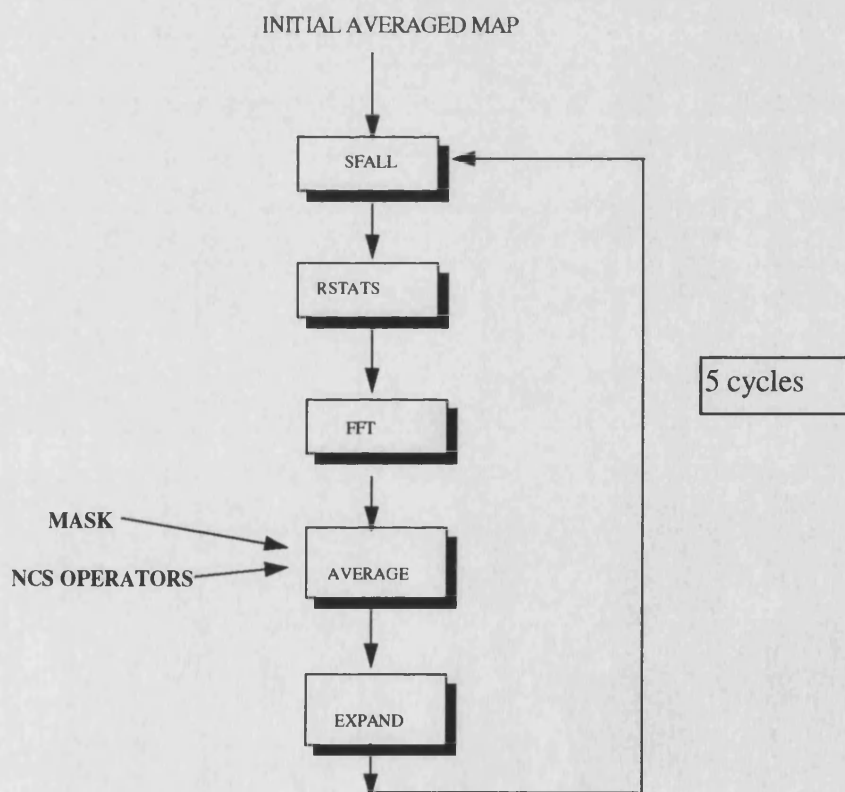


Fig. 5.4 Representation of the RAVE scheme.

The mask

The ncs operators had been defined, the SIR map had been calculated; to fulfil the averaging criteria a mask needed to be generated. The mask is a 3 dimensional grid of 0's and 1's, with all pixel points within the mask set to 1, and all points outside the mask set to zero. The mask can be generated from a set of atomic co-ordinates or BONES skeleton points, whereby a certain radius is assigned for each atom, which consequently governs the size of the mask. Once the initial mask has been constructed, it can be manipulated in various ways (using MAMA): trimming overlapping edges of symmetry related masks, expanding, contracting, removing voids within the mask and removing extraneous mask points. Additionally, the mask can be edited using the commands within the graphics 'O' package.

However, because GDH had no known structural homologue, a mask could not be generated from a set of co-ordinates. Furthermore, although the SIR map and the corresponding BONES representation allowed an approximate delineation of the tetramer boundary, it was decided that a mask could not be confidently constructed without extensive editing of the highly ambiguous set of bones. To overcome the hurdle of not having a mask, a spherical mask of 43Å radius was constructed, centred at the origin of the 222ncs (the centre of mass of the tetramer). A radius of 43Å was chosen so that the symmetry related sphere would not overlap significantly. Any overlapping areas were removed using MAMA. Obviously this crude spherical mask was insufficient in defining the boundary of the GDH tetramer, however use of this mask would indicate whether the averaging procedure had any potential to improve the quality of the phases of the initial SIR map (fig. 5.5 & 5.6).

Averaging within the sphere

Using the spherical mask, 3 different averaging rounds were completed. After the first round of averaging, the low resolution terms were included and then the ncs operators were refined and used in subsequent averaging (using the automatic ncs improvement option of IMP).

	Cycle	1	2	3	4	5
ROUND 1	Reliability index	0.408	0.374	0.388	0.326	0.319
	Mean correlation coefficient	0.595	0.746	0.796	0.815	0.825
ROUND 2	Reliability index	0.411	0.304	0.266	0.252	0.245
	Mean correlation coefficient	0.632	0.781	0.814	0.834	0.840
ROUND 3	Reliability index	0.411	0.310	0.260	0.241	0.233
	Mean correlation coefficient	0.632	0.783	0.827	0.849	0.855

Table 5.2 **ROUND 1:** GRID: 54 80 60, Mask: spherical mask, resolution limits: 10 -4.5Å, unrefined symmetry operators. **ROUND 2:** GRID: 54 80 60, mask: spherical mask resolution limits: 50 -4.5Å, unrefined ncs operators. **ROUND 3:** GRID: 54 80 60, mask: spherical mask, resolution limits: 50 -4.5Å, refined ncs operators.

IMP optimises the correlation coefficients between the density inside the mask and density within that mask after application of the ncs operator. The refined symmetry operators were:

$$\mathbf{A \text{ to B: }} \begin{Bmatrix} -0.824 & -0.484 & 0.295 \\ -0.484 & 0.331 & -0.810 \\ 0.294 & -0.810 & -0.507 \end{Bmatrix} \{28.426 \quad 92.922 \quad 135.699\}$$

$$\mathbf{A \text{ to C:}} \left\{ \begin{array}{ccc} -0.704 & -0.280 & -0.653 \\ -0.285 & -0.731 & 0.620 \\ -0.651 & 0.622 & 0.435 \end{array} \right\} \{70.818 \quad 71.385 \quad 1.259\}$$

$$\mathbf{A \text{ to D:}} \left\{ \begin{array}{ccc} 0.527 & 0.769 & 0.362 \\ 0.769 & -0.613 & 0.183 \\ 0.363 & 0.181 & -0.914 \end{array} \right\} \{-64.061 \quad 80.945 \quad 98.400\}$$

The importance of including the low resolution terms in the initial reflection file and a more accurate description of the ncs operators, is indicated by the improved averaging statistics. The resultant averaged map(s) showed clear 222ncs with interpretable features. Inspection of the map for secondary structural elements revealed an area of density suggestive of 12 strands of β -sheet (fig. 5.7), with one of the ncs 2-folds passing through the centre. The nucleotide-binding domain contains a β -sheet of 6 strands, with the LADH monomer dimerising via these domains (to give a 12 stranded β -sheet). Superposition of the α carbon backbone of the LADH dimer onto the averaged density revealed a striking correlation between the density and the LADH α carbon backbone. In addition some of the flanking helices present in the LADH nucleotide binding domain fitted regions of the observed electron density, although variations in the position and length of these flanking helices were evident. The only conserved sequence motif in the nucleotide binding domain is the GxGxxG/A motif. Alignment of this LADH region (β A-turn- α A) to that of the observed density again revealed good correlation (fig. 5.8). These observations strongly indicated that GDH contained the nucleotide binding domain and furthermore shared striking similarities to that of the nucleotide binding domain of LADH, especially the mode of subunit/subunit interaction.

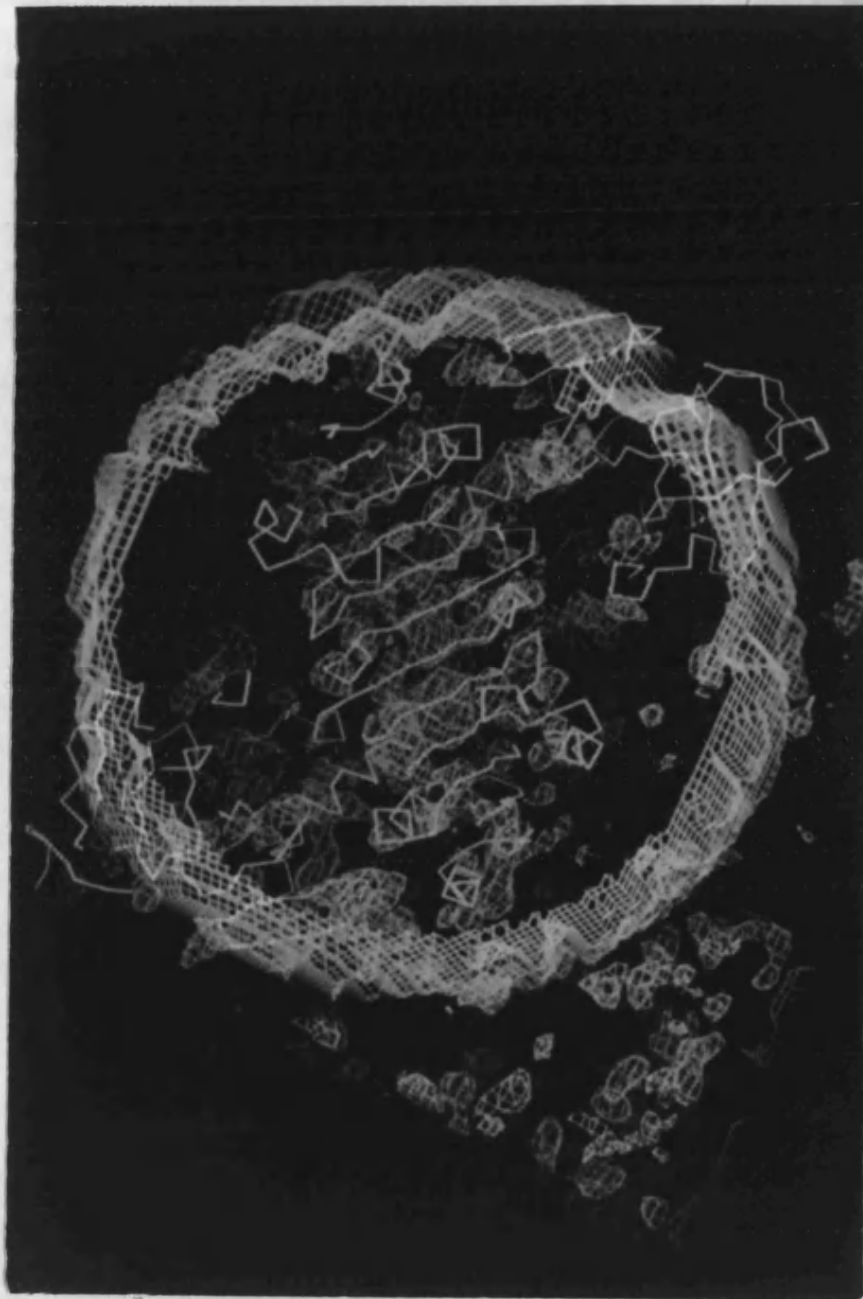


Fig. 5.5. The spherical mask and the GDH SIR density.

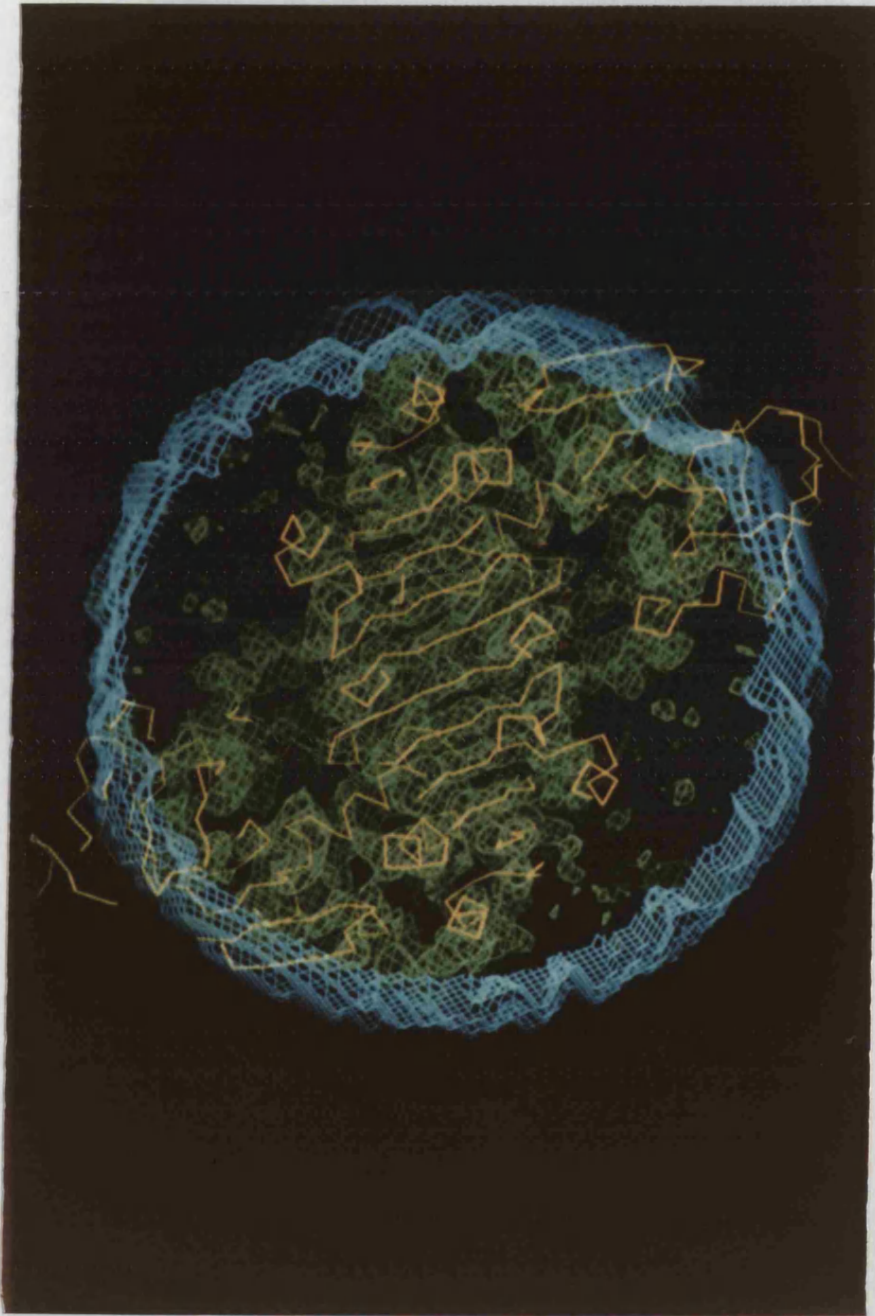


Fig. 5.6. The corresponding GDH density after averaging within the sphere. The 2-fold symmetry can be visualised.

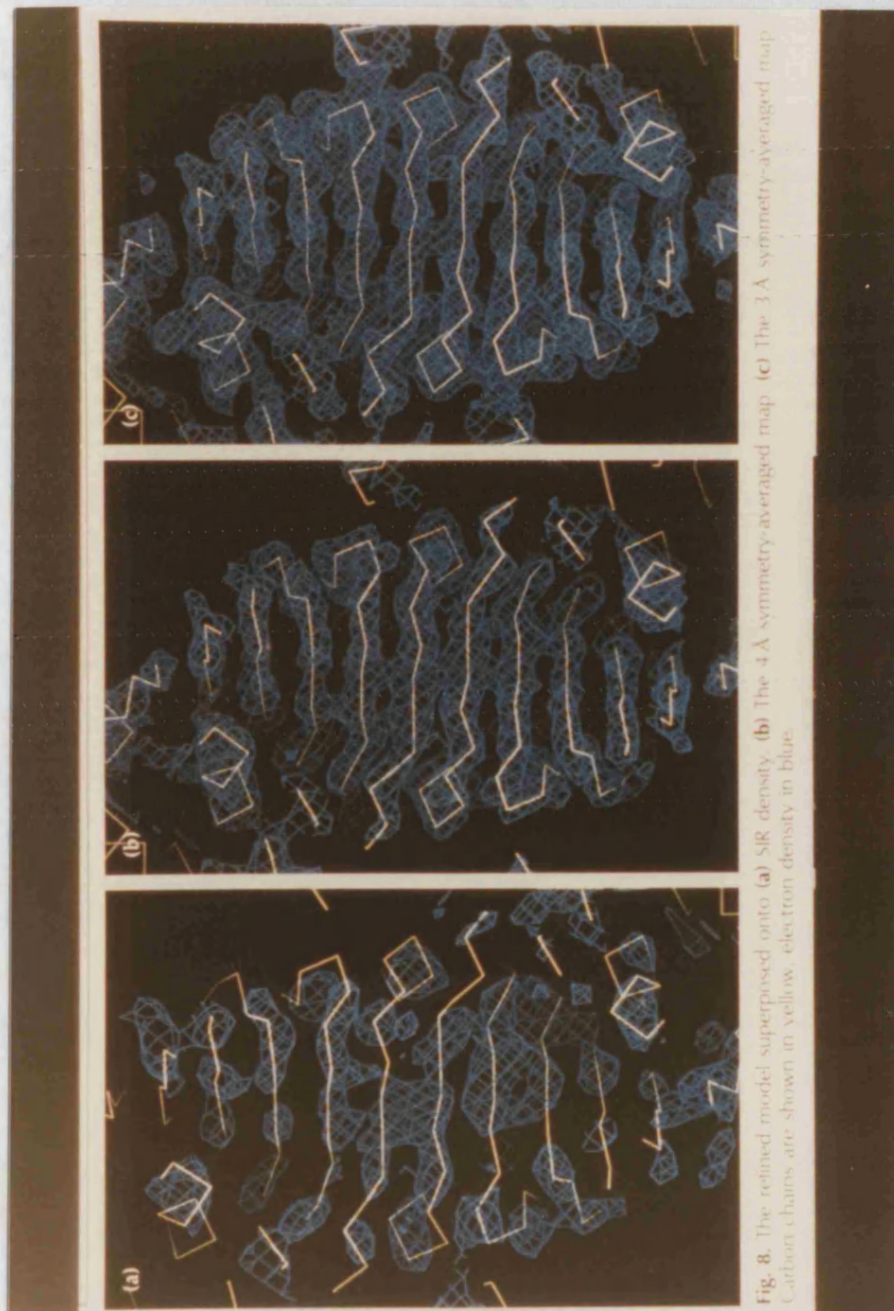


Fig.5.7 The refined model superposed onto a) SIR density. b) The 4Å symmetry averaged map. c) The 3Å symmetry averaged map.

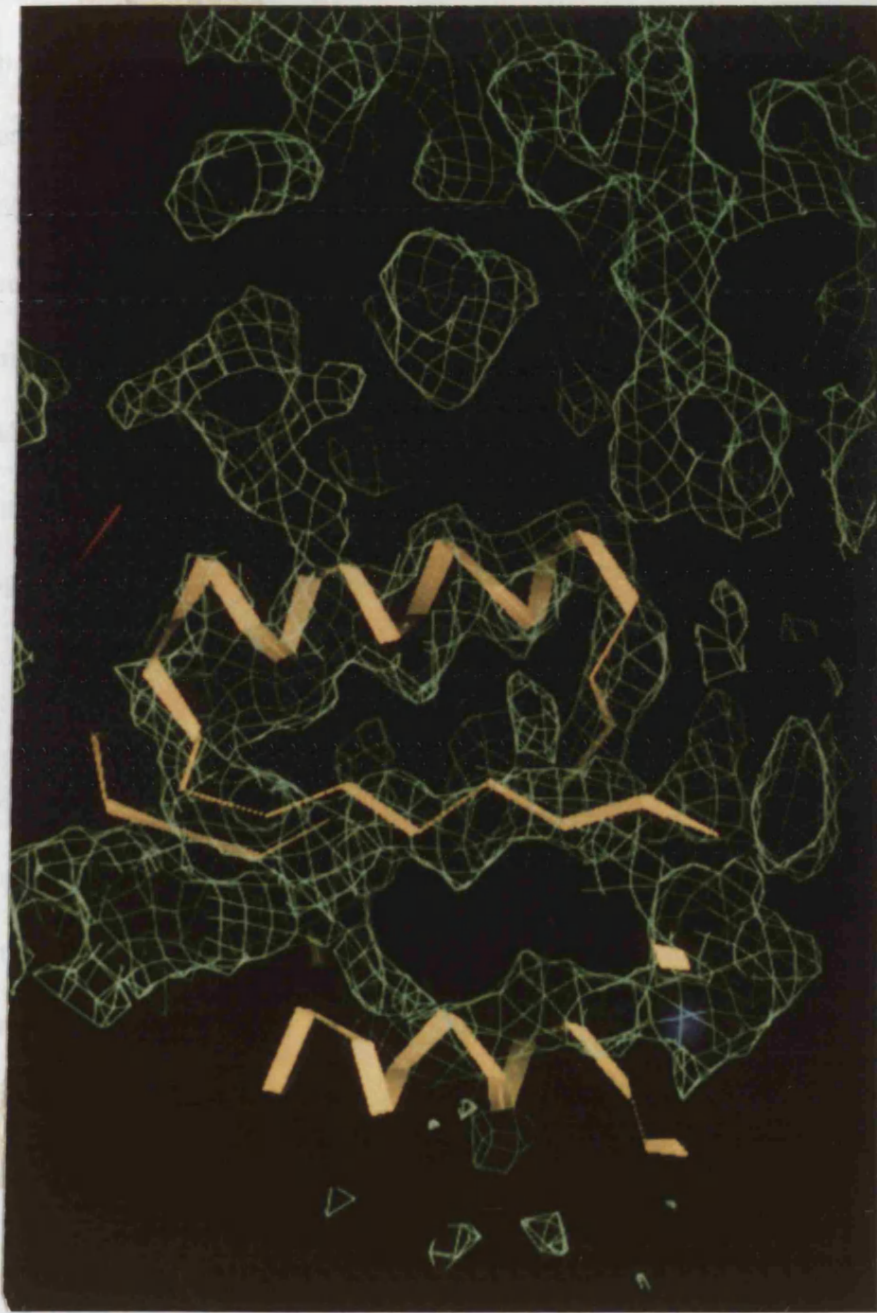


Fig. 5.8. The α -carbon backbone representing the GxGxxG region of LADH superposed onto the corresponding area of GDH density.

The Alcohol dehydrogenase mask

Following the apparent success of the averaging, the map was extended to 4Å resolution using the 4Å SIR phases, with a concomitant change of the mask grid (GRID 60 96 66). The final mean correlation coefficient for the ncs operators was 0.8 and the reliability index was 0.23.

The averaged map appeared to be constrained by the spherical mask. The apparent structural similarity between GDH and LADH in the nucleotide binding domain, and the tentative sequence relationship between GDH and the long-chain alcohol dehydrogenase superfamily (Zn-ADH), suggested a similarity in their overall three-dimensional fold.

GDH is a tetramer, whereas LADH is a dimer. There are no reported structures of a tetrameric alcohol dehydrogenase, although a model of the tetrameric yeast alcohol dehydrogenase (YADH) had been constructed from LADH. However, no definitive explanation was proposed concerning the tetrameric association of YADH. A tetrameric model of the LADH dimer was constructed:- the ncs operator (A to B) was applied to the co-ordinates of the dimer that had been superposed onto the density of GDH. A mask was then constructed around the co-ordinates of the modelled 'dimer of dimers' tetramer. Graphics visualisation of the resultant mask revealed that the mask fitted the LADH co-ordinates too tightly, and exhibited a large void at the centre of the 222 ncs. The mask was expanded accordingly, the overlapping edges of the symmetry related mask trimmed, and the void partially filled (fig. 5.9). This mask was used for the next round of averaging (ROUND 4, table 5.3) (fig. 5.10):

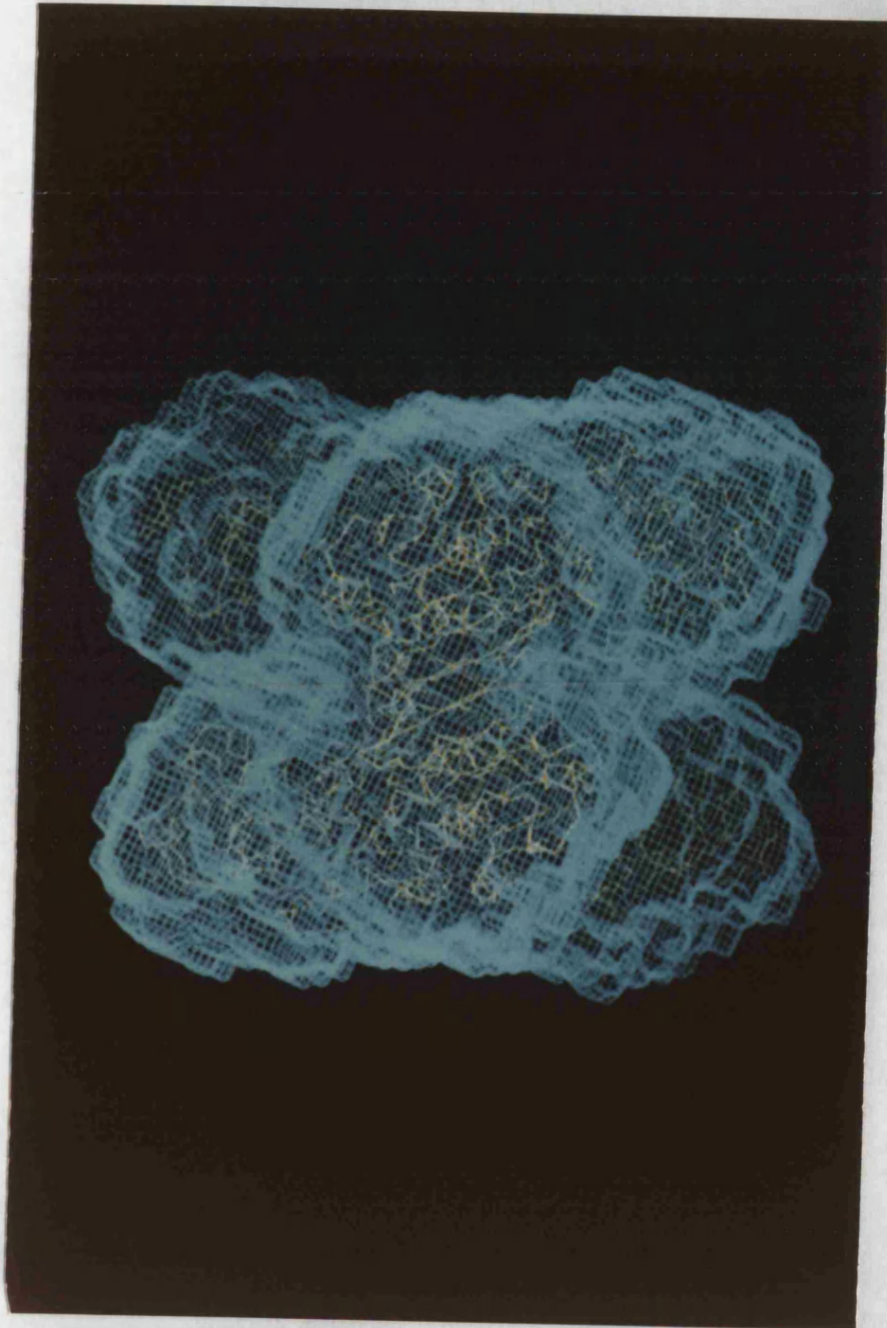


Fig. 5.9. Global representation of the LADH tetrameric mask.

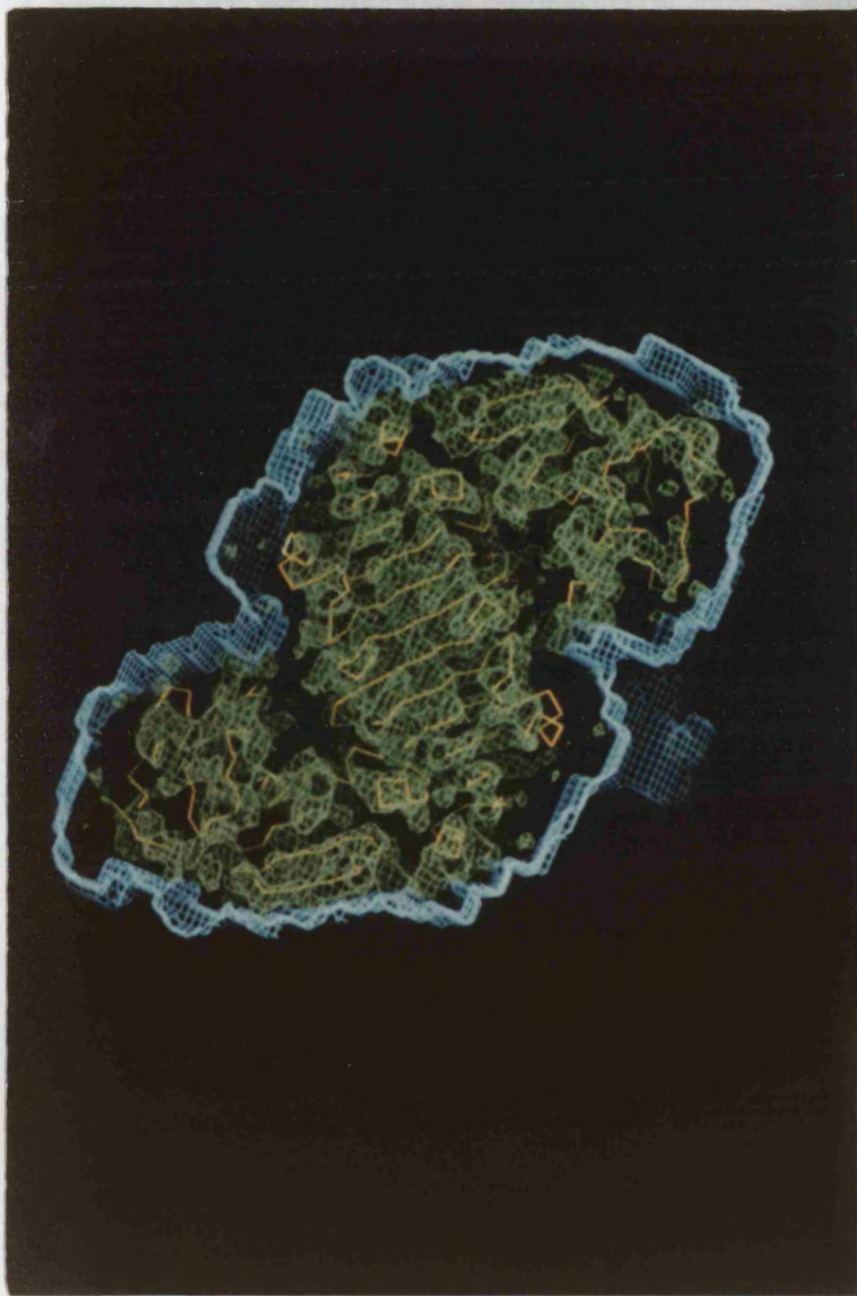


Fig. 5.10 Averaged density within the LADH tetrameric mask.

ROUND 4	Cycle	1	2	3	4	5
	Reliability index	0.408	0.270	0.224	0.198	0.191
Mean correlation coefficient	0.59	0.71	0.74	0.77	0.79	
ROUND 5	Reliability index	0.445	0.298	0.219	0.190	0.180
	Mean correlation coefficient	0.55	0.73	0.79	0.82	0.83

Table 5.3 ROUND 4 and ROUND 5 averaging statistics at 4Å resolution using the ADH tetrameric masks

In round 4, the correlation coefficients of the ncs operators decreased slightly, however the ncs reliability index was improved significantly, indicating the validity of using the tetrameric mask.

The current LADH mask was not optimal; there were regions where the LADH tetramer protruded out of the mask, and the void at the centre of the 222 ncs was still prominent. The mask was expanded twice, the voids filled, the overlapping edges trimmed twice, and finally contracted to produce the optimal ADH mask at the working resolution of 4Å. Using the optimal ADH mask the averaging statistics were improved (ROUND 5, table 5.3):

Phase extension

Generally, the native data set frequently extends to a higher resolution compared to the data sets of the derivative(s). The phases can be gradually extended towards the resolution limits of the native data by various methods. Typically at 3.5Å resolution,

side chain density of the larger amino acids can be discerned. To interpret the averaged map and fit the primary sequence of GDH to the density, the phases were extended to higher resolution. The final data set containing the 2.9Å data had not been collected; the current optimum data set only extended to 3.5Å.

Two modes of phase extension to 3.5Å were undertaken:

- 1) Phase extension using SQUASH, and subsequent averaging using RAVE.
- 2) Concomitant phase extension and averaging in RAVE.

The SQUASH phase extension statistics were promising, however the resultant averaged map was very disjointed.

Using the second method, the phases were gradually extended in steps of 0.01Å prior to each round of averaging. A 0.01Å phase extension resulted in approximately 100-200 native higher resolution reflections being included in the averaging round. As the grid is a function of the resolution of the data, the grid for averaging and the mask, was set for 3.5Å (GRID: 72 104 76)

ROUND 6	initial	final
reliability index	0.160	0.159
correlation coefficients	0.776	0.769
	0.764	0.759
	0.800	0.797

Table 5.4 Averaging statistics for ROUND 6.

The averaging statistics during ROUND 6 did not change considerably during the phase extension, but the resultant 3.5Å map displayed greater definition of side chain density, enabling model building into the density to begin (chapter 6)

The GDH mask (mark 1)

Once a sufficient degree of the GDH carbon backbone of the monomer had been traced, the tetramer was generated in `O` using the ncs operators. A mask was constructed around the tetramer co-ordinates, and manipulated accordingly (MAMA) to produce an optimum mask.

Concurrently the 2.9Å native data was obtained. Averaging within this new GDH mask and phase extension to 3Å (in 0.005Å steps) using RAVE produced a map with a marked improvement in side chain definition, and enabled some of the surface loops to be traced more readily

The GDH mask (mark 2)

Further model building into the density, including the assignment of side chain residues, indicated a problem in the refinement (Chapter 6, round 4). There was a large break in the density at the surface of the molecule. The LADH and sorbitol dehydrogenase (SDH) co-ordinates were superposed onto the GDH model. A large surface loop in LADH (and SDH) (residues 79-85) was present at the corresponding area of the GDH truncated density, and furthermore this loop protruded beyond the limits of the mask.

The GDH mask was edited in `O` to include this additional loop : a 6Å box was constructed around these loops (fig. 5.11) and merged with the existing mask (fig. 5.12 & 5.13). Averaging within this new mask (mark 2) (final reliability index = 0.1567, correlation coefficients (0.812, 0.811, 0.827), with simultaneous phase extension from 4Å to 3Å, finally produced a map in which the nucleotide binding domain and the catalytic domain could be interpreted. (see chapter 6).

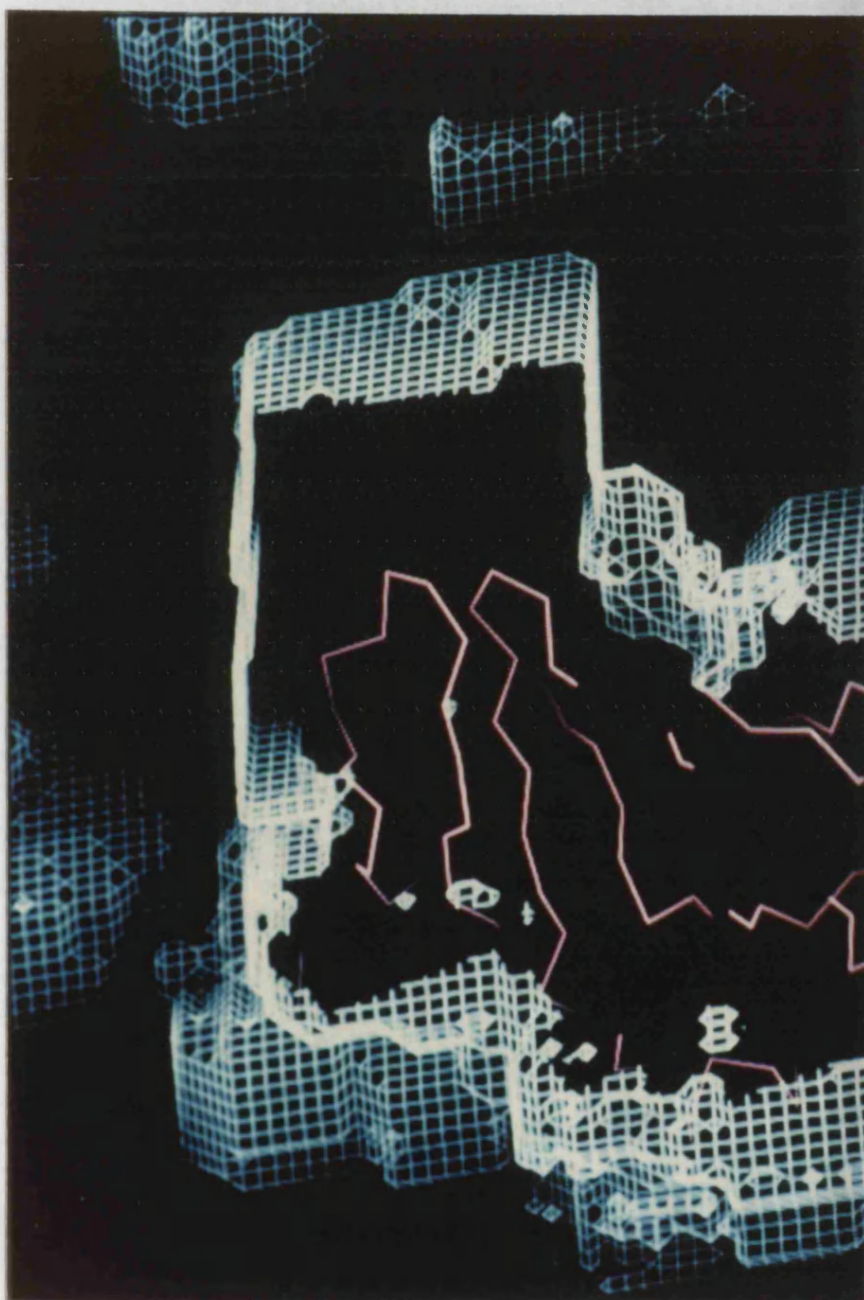


Fig. 5.11. The annex to the GDH mark 2 mask. LADH α carbon backbone is in magenta.

Fig. 5.12. The GDH tetrameric mask containing the refined GDH tetramer. Each monomer is represented as a separate colour.

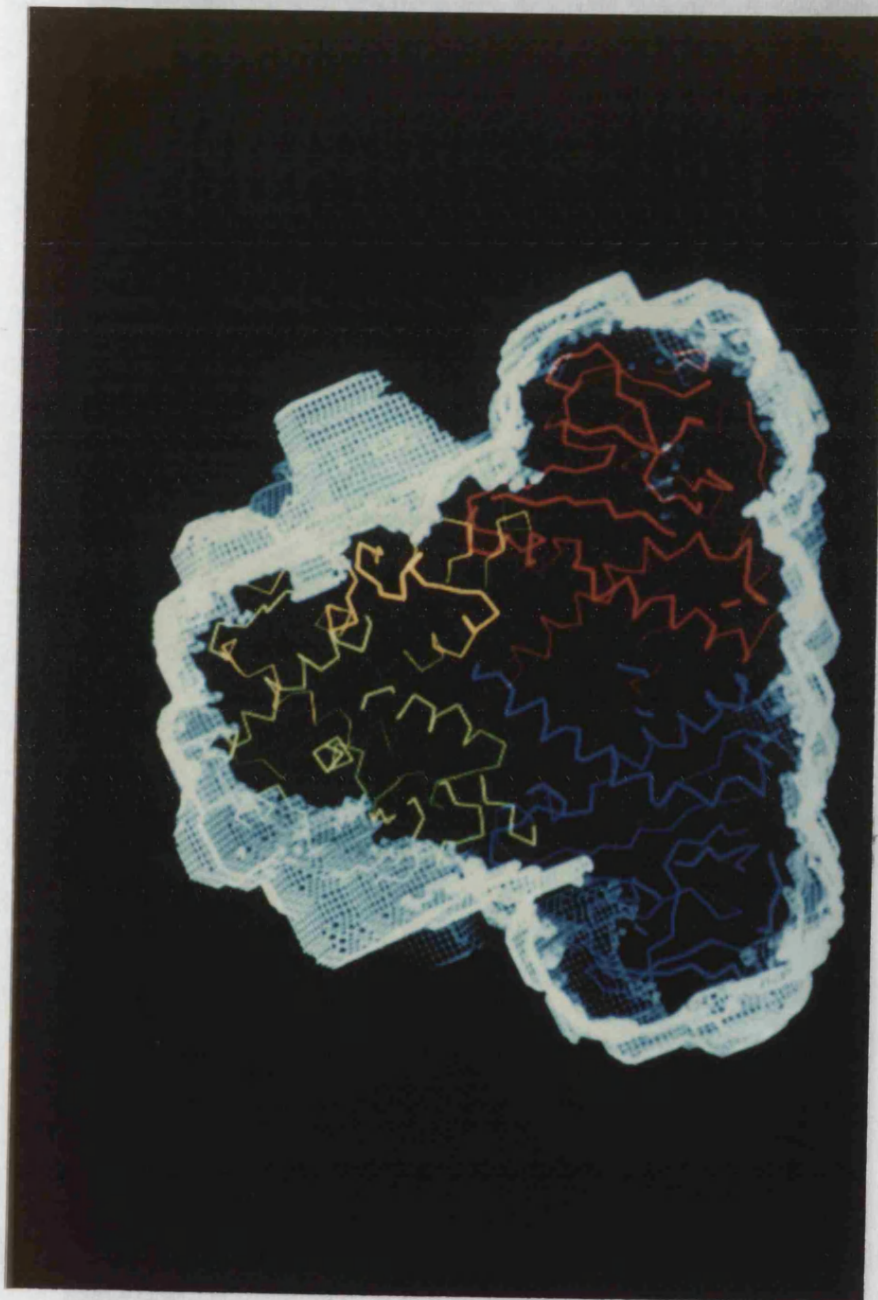


Fig. 5.12. The GDH tetrameric mask containing the refined GDH tetramer. Each monomer is represented as a separate colour.

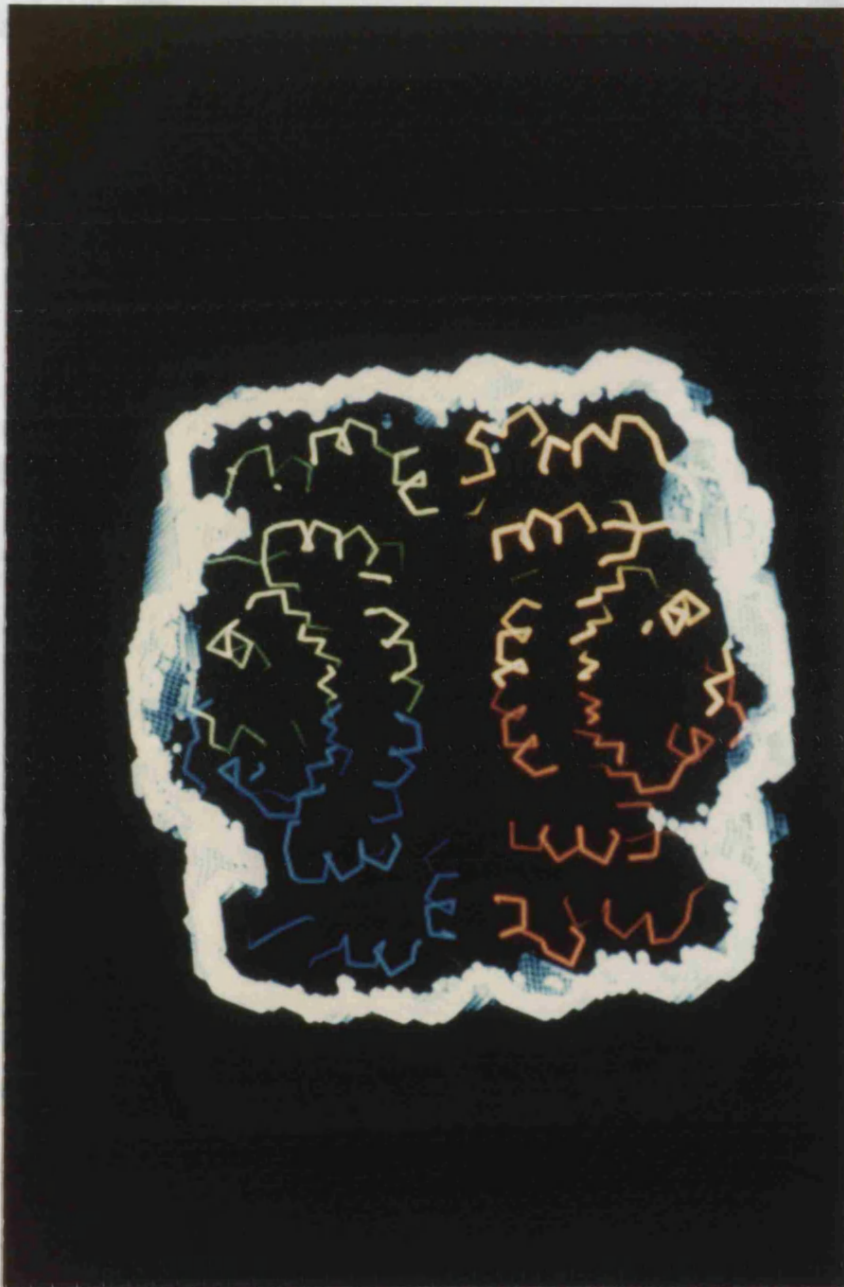


Fig. 5.13. Another view of the GDH tetramer contained within the tetrameric mask (mark 2).

Chapter 6

MODEL BUILDING AND REFINEMENT

Introduction

The quality and resolution of the X-ray structure factor amplitudes and the accuracy of their associated phases are parameters that contribute to the quality of the electron density map and hence determine the ease of map interpretation. Map interpretation is probably the most subjective area in crystallography, and a wrongly built model may be difficult to correct. The stages in model building are :

- Identify the molecular boundaries.
- Produce a continuous skeletal representation of the density.
- Assigning the C α positions of the polypeptide chain.
- Fit the primary structure to the density.

All the map interpretation and subsequent model building was performed using the 'O' program (Jones *et al.*, 1991) on an Evans and Sutherland ESV10/33. Model building was first attempted using a 3.5Å solvent flattened, averaged map. Inspection of the map had revealed secondary structural elements, including the 12 stranded β -pleated sheet. This observation led to the eventual development of a mask that enveloped the entire tetramer. There were two stages in the model building of the GDH structure:

defining the nucleotide-binding domain and tracing the catalytic domain. The former domain was relatively straightforward - the density was very continuous and the LADH nucleotide-binding domain proved to be a useful guide. The catalytic domain proved more difficult; primarily because there was no apparent homologue for the catalytic domain. Subsequent analysis also revealed that some of the surface loops of the catalytic domain were truncated by the initial mask that were used in averaging and that the α carbon backbone of GDH was initially built through the electron density occupied by the zinc ions. Obtaining the 3Å averaged map for the complete tetramer eventually enabled the GDH structure to be satisfactorily built and refined.

The nucleotide-binding domain is a very conserved feature amongst a number of dehydrogenases, in particular the GxGxxG/A loop of the $\beta\alpha\beta$ fold. This region served as a site for initial model building into the density, with further map interpretation extending either side to this region.

Topology determination

The connectivity of the GDH density was initially estimated using BONES in conjunction with 'O' (O_bones)(Jones *et al.* 1991). Bones is used to produce a skeletonised representation of the electron density, and analyses the degree of connectivity. There are a number of levels of bones used within O: for GDH, probable main chain, possible main chain and side chain were represented by levels one, two and three respectively, with each level being colour coded. The degree of skeleton connectivity reflects the connectivity of the electron density map. As a consequence of the phase constraints imposed by the averaging procedure, the GDH map had a high degree of connectivity (see fig. 6.1), which aided in the eventual

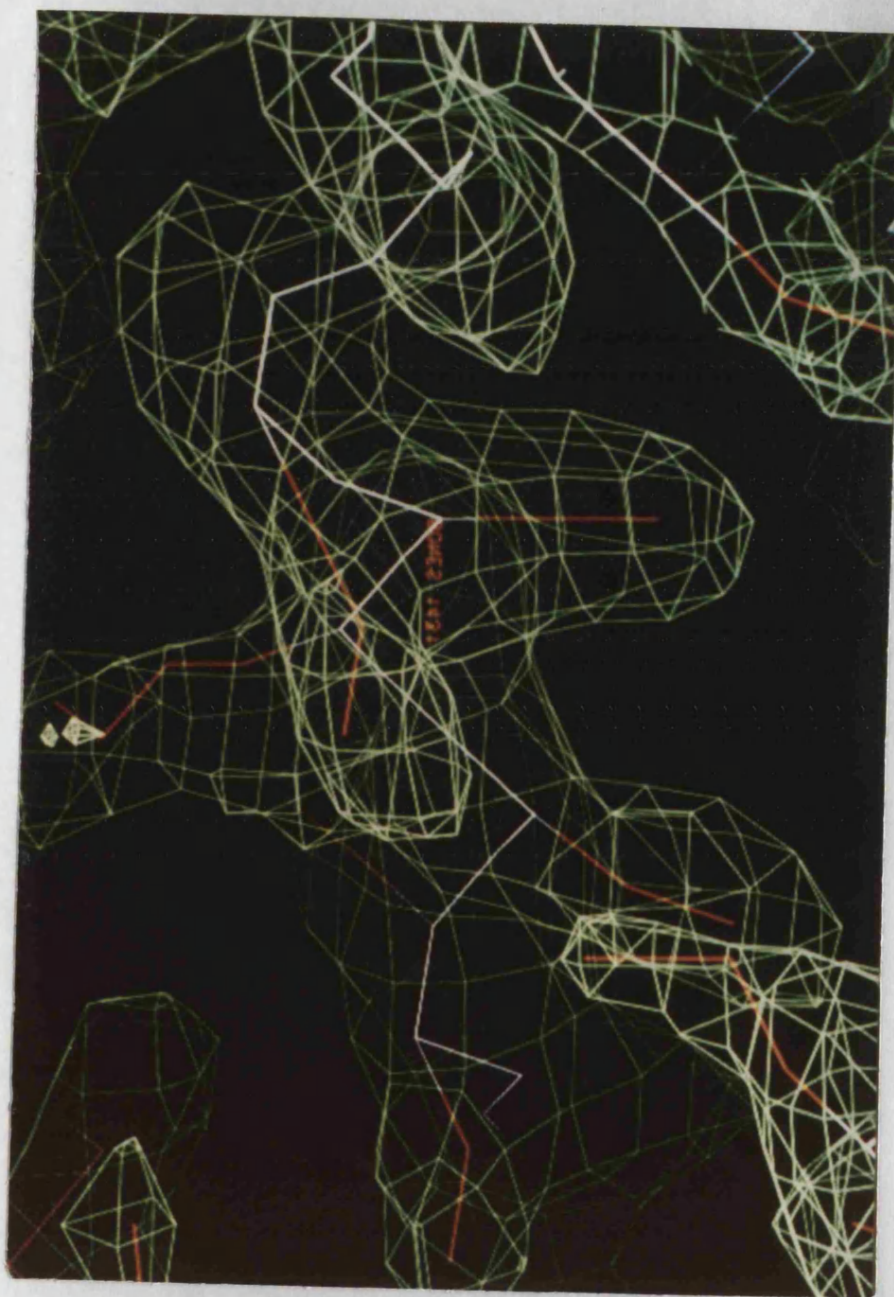


Fig. 6.1. Bones representation of a helical region of the GDH density.

topology determination. The skeletonised representation of the averaged map contained one long, continuous strand of possible main chain. However, inspection of areas of the averaged map still revealed errors in the connectivity of the corresponding skeletal atoms. Where the connectivity of the skeletal atoms were deemed incorrect, the *bones_break* and *bones_make* commands allowed the skeleton to be manipulated accordingly to produce the desired connectivity. Once the connectivity of an area of density had been satisfactorily assessed, the possible main chain was designated probable main chain. Starting at the GxGxxG/A region and working upstream, bones skeletons that adequately described the fold of the chain were assigned probable main chain. Producing a continuous skeletal strand within the nucleotide-binding domain was aided by superposition of the α carbon backbone of LADH. The LADH α carbon backbone deviated from the GDH density mainly at the loop regions and for some helices of the putative nucleotide-binding domain of GDH. Nevertheless at these deviant regions, a clear path of connectivity for GDH could be visualised, the bones atoms edited where necessary, and assigned to probable main chain.

Once the connectivity of the nucleotide-binding domain was complete, map interpretation journeyed into the catalytic domains. Immediately C-terminal to the Rossmann fold was a helical region of density, in which the bones atoms were readily edited to produce the desired connectivity. However upstream of this helix, breaks in the density were encountered, which made assignment of main chain difficult.

Attention then turned to the N-terminal catalytic domain. Within this domain, bones editing primarily concentrated on obvious areas of density that represented secondary structural elements. The connectivity between these secondary structural elements were naturally more difficult to assess.

There were two areas of knotted density (the future zinc sites) that were particularly hazardous for determining the fold of the protein, because both these regions appeared to form closed loop structures. In addition, breaks in the density, which mainly occurred at the surface of the N-terminal domain, were more prevalent. In these areas only a dubious guestimate of the connectivity could be made (after long hours of staring at the graphics). When the going became too difficult, attention switched to the next stage of model building.

Assignment of C α positions

Starting from the GxGxxA region, and using the LADH nucleotide-binding domain as a guide, the probable main chain bones atoms were moved (*bones_move*) adjacent to areas that represented side chain density. The *bone_pick_C α* command then transformed the bone atom to a putative C α atom. Various conformational factors, guided the choice of C α positions, such as the C α -C α bond is approximately 3.5Å, there are 3.6 residues per helical turn, and adjacent residues point out on alternate sides of a strand. Once the C α positions had been assigned for the nucleotide-binding domain, and a large proportion of the catalytic domain (with approximately 20 amino acids missing from both termini), the initial bad contacts of the polyalanine model were removed (*auto_main_chain*) in 'O'. The model was then refined in X-PLOR.

Fitting the primary sequence

The primary sequence of GDH (Bright *et al.*, 1993) could only start to be confidently fitted to the density when the 3Å map was obtained, although attempts were initially

made to fit the sequence to the 3.5Å map. Starting from the GxGxxA loop, sidechains were appended to the polyalanine backbone (*mutate_replace*), and a rotamer of the particular residue that optimally fitted the observed density was chosen from the 'O' database (*lego_side_chain*). After the nucleotide-binding domain was traced, fitting of the primary sequence ventured into the catalytic domain, where tracing of the chain proved to be more difficult. Several rounds of refinement and map calculations were necessary before the complete catalytic domain was accounted for.

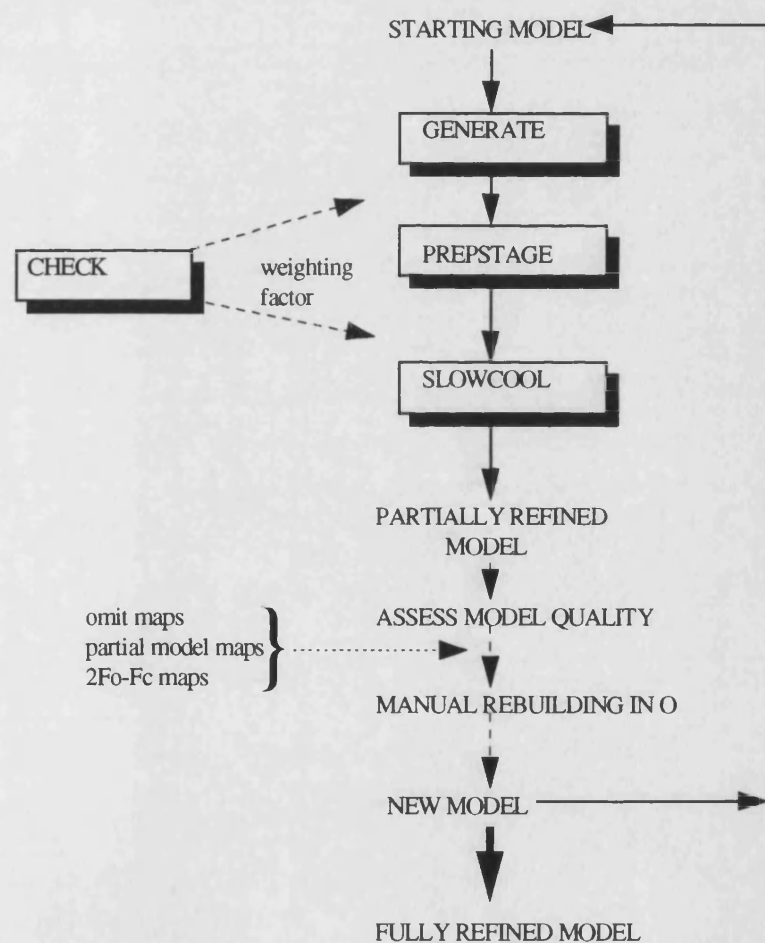


Fig. 6.2 Schematic representation of X-PLOR and the refinement stages.

Refinement

Refinement is a procedure in which an initial model structure is modified to give improved agreement between observed and calculated structure factor amplitudes, while imposing acceptable model stereochemistry. It can be represented as a minimisation of E_{tot} where :

$$E_{tot} = E_{xray} + E_{stereochem}$$

$$E_{xray} = \sum w (|F_o| - |F_c|)^2, \text{ a pseudo-energy term.}$$

$$E_{stereochem} = \sum (E_{bond} + E_{angles} + E_{planes} + E_{chiral} + E_{torsion} + E_{vdw} + E_{electrostatics})$$

There are many refinement packages available for protein structure refinement (for example, X-PLOR (Brünger, 1988) and TNT (Tonrud *et al.*, 1987)). The X-PLOR (version 2.1) refinement procedure (see fig. 6.2) was used in the refinement of the GDH structure. GENERATE: generates the oxygen and nitrogen hydrogens of the model prior to refinement, and prepares a topology file describing the bonds, angles *etc* in the protein. For conventional least squares minimisation (PREPSTAGE) and the molecular dynamics minimisation (SLOWCOOL), a weighting factor for the X-ray term is required, which is determined by CHECK. CHECK calculates the weighting factor by carrying out a short dynamic simulation in the absence of the X-ray terms , and calculates a weight which balances the derivatives of the X-ray and stereochemical terms. The weights (an estimate of the bond strength, bond angle *etc*) are also required for the different stereochemical terms, which are derived from the study of small molecules, or spectroscopy. PREPSTAGE, conventional least squares minimisation is used to reduce the initial bad strain and contacts within the model. Due to the limited radius of convergence of this method, a local energy minimum is

achieved, as opposed to the apparent global minimum found in SLOWCOOL. In this approach, kinetic energy is given to the system by assigning atomic velocities from a Boltzman distribution corresponding to a temperature of 2000K, to overcome any potential energy barriers between the local and global free energy minimum. The system is then gradually cooled down to yield the minimised structure.

A measure of the agreement between F_o and F_c is given by the R-factor :

$$R = \frac{\sum ||F_{OBS}| - |F_{CALC}||}{\sum |F_{OBS}|}$$

The R-factor for a well refined protein structure should be below 20%. After a round of refinement, the stereochemical quality of the (partially) refined structure can be assessed using PROCHECK (Morris *et al.*, 1992). The goal of refinement is to try and reduce the R-factor and maximise the stereochemical quality of the structure.

NCS restraints

For the initial rounds of refinement the *ncs strict* card in X-PLOR was implemented, which treated each monomer in the tetramer as being identical. All the energy calculations were solely performed on the monomer. The ncs operators from RAVE were used to generate the other monomers of the tetramer.

In the latter stages of refinement, the ncs strict card was removed and an ncs restraint was applied. Here the ncs related atoms(x) were restrained in their average positions by the high energy term:

$$E_{NCS} = w(x - \bar{x})^2$$

\bar{x} = average position

w = weight (300 Kcal/mole/Å²)

This permitted slight differences between the monomers. Eventually, when the majority of the refinement was completed, the ncs card was removed altogether.

Temperature factor refinement

The decision of how to refine the temperature factors (B factor) depends strongly on the number of observations / parameter ratio, *i.e.* the resolution of the data. Isotropic B factors can only be refined at a resolution greater than 2.7Å, with reasonably complete data, whereas to refine anisotropic B-factors requires very high resolution. In the initial stages of refinement, the B values were initially fixed at 20Å². In the latter stages of refinement, because the resolution of the GDH data only extended to 2.9Å, even isotropic refinement was not valid - the B factors had to be grouped to lower the parameter/ observation ratio. Two groups were selected for each residue : the backbone and the side chain atoms with uniform shifts being applied to each group. Subsequently an overall B factor for the tetramer was also determined.

Refinement round 1

See table 6.1 for summary of refinement statistics. The polyalanine tetramer was generated from the monomer in 'O' using the ncs operators. The model was refined with the ncs strict card. The R-factor dropped approximately 15% (from 49.3% to 35.4 %), which was an encouraging sign for a partially built polyalanine model.

A rigid body refinement (10→3Å) of the nucleotide-binding domain (with ncs strict) was undertaken with the aim to improve the ncs operators. A refinement using these 'X-PLOR ncs operators' did not decrease the final R-factor - indicating that the ncs operators were optimally defined in RAVE.

A mask was created in RAVE (mark 1) using this refined polyalanine tetrameric model (chapter 6).

Rounds 2 + 3

Acquisition of the phase extended 3Å map enabled the nucleotide-binding domain to be traced. Incorrectly assigned residues were changed (*mutate_replace*). In the cases where there were an inappropriate number of residues built in a helical/turn region, the *mutate_delete* or *mutate_insert* commands were used. Following this, the *lego_loop* command enabled a reassessment of the backbone conformation of that area.

The traced nucleotide-binding domain served as a template for superposition (SHP) of the LADH monomer. From inspection of this superposition, the 2 zinc sites of LADH corresponded well with 2 regions of high electron density ($>10\sigma$) in the GDH map, indicating the possibility of GDH containing zinc. However, the C α backbone of GDH had been built through these regions of high density, and had to be revised. The chain was traced from the nucleotide-binding domain to the first zinc site (the putative structural lobe), which indicated that 3 cysteines were located in this vicinity, with one of the cysteines being confidently assigned as a ligand to the zinc. The remaining residues within this lobe stayed as polyalanine. The resultant model refined to an R-factor of 30.4% (Round 2).

The chain was traced beyond the first zinc site towards the second zinc site.

However, a large break in the density was encountered, and the putative cysteine and glutamate ligands were distant from the second zinc site. The first 80 N-terminal residues retained their polyalanine status. This new model refined to 29.6% (Round 3).

Round 4

Biochemical evidence (see chapter 7) suggested that GDH may be structurally more homologous to SDH than LADH. Superposition of the SDH model onto the partially refined GDH model revealed this to be true (see chapter 8). SDH and LADH possessed a surface loop which corresponded to the large break in electron density of GDH. Furthermore, this SDH and LADH loop (approximately 18 residues) protruded out of the current mask (mark 1, chapter 5), as did another smaller surface SDH loop. It appeared that the mask restricted the GDH model at these surface loop positions, so the mask was edited in 'O' to incorporate these LADH and SDH loops. Subsequent averaging within this new mask (mark 2, chapter 5) revealed that GDH also possessed a similar surface loop. The newly averaged map enabled the surface loop to be traced, and resulted in the putative zinc ligands (Cys 40, Glu 67) falling in the proximity of the putative catalytic zinc. In addition, this new map enabled the topology of the N-terminal region of the catalytic domain to be confidently delineated and subsequently enabled the majority of the remaining residues (approximately 60) to be traced. The new model was refined to 28.0% (after slowcool) and 25.4% after grouped B-factor refinement.

Round 5

The refined model from round 4 was assessed using PROCHECK (Morris *et al.*, 1992) (see assessment of model quality). The residues that displayed bad geometry and high temperature factors were highlighted - subsequent refinement was focused on these erroneous areas. 2Fo - Fc maps, partial model maps served as an aid in the model rebuilding. The flow chart (fig. 6.3) depicts the general scheme for creating 2Fo-Fc maps and partial model maps. SFALL, RSTATS and FFT have been described in previous chapters (5 & 6). SIGMAA calculates weighted Fourier coefficients (FWT) from the calculated phases of a partial structure combined with the observed phases. This serves to improve the quality of the phases and hence the quality of the map. However these maps are biased towards the model, and an incorrectly built model would not easily be detected. For uncertain regions of the structure, an OMIT map was calculated (in X-PLOR) where the specified region was excluded from the phase calculation. This reduces the bias from the model in these uncertain areas. Inspection of the resultant OMIT map may indicate the true fold of the protein in that region. Omit maps were calculated around the structural lobe and the region C-terminal to the sole tryptophan. In conjunction with the 2Fo-Fc maps, the structural lobe was traced (with the cysteines co-ordinating to the putative zinc). *Lego_loop* and *lego_mainchain* of stereochemically inadequate regions allowed alternative peptide conformations to be explored. The tryptophan region caused more of a problem. An alternative conformation of the C-terminus was explored, the R-factor increased after subsequent refinement of that model, and the 3D-1D score (Luthy *et al.*, 1992)(see model quality) was negative for that region. Eventually a satisfactory conformation was found. The model was refined with relaxed ncs restraints (to allow slight structural deviation

between monomers). The model refined to 20.8% after grouped B-factor refinement (25% after slowcool).

Rounds 6 + 7 + 8

Model building was focused on the Ramachandran outcasts. The ncs restraints were removed, and the model refined to 18.1% (20.4% after slowcool). Further model

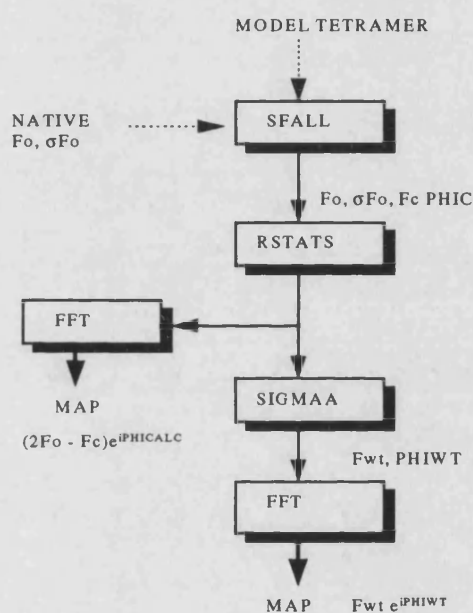


Fig. 6.3 Flow chart depicting the scheme to create partial model maps and 2Fo-Fc maps.

building in the refined structure eliminated more disallowed residues, giving a model which refined to an R-factor of 17.9% (with a 2σ cut-off).

Removal of the ncs restraints produced no appreciable deviation of the main chain of the monomers in the tetramer compared with that of the restrained model (root mean square difference between restrained and unrestrained monomer is 0.66\AA for all C_α atoms). For all reflections between 8 and 2.9\AA , the model refined to 19.3%.

Additionally, the model refined to an R-factor of 20.5% when an overall temperature factor refinement with individual atomic B-factors set to 20\AA^2 was implemented.

There were 3400 atoms/monomer, and a total of 29000 unique reflections measured to 2.9\AA . Using the overall temperature factor refinement, the observation/parameter ratio was 1:2, whereas using grouped B-factor refinement, the ratio was 1:1.

Round	R-factor (%)	NCS	Resolution (\AA)	Temperature factor
1	35.4	Strict	3.5	none
2	30.4	Strict	3.0	none
3	29.6	Strict	3.0	none
4	25.4	Strict	3.0	b-group
5	20.8	Relaxed	3.0	b-group
6	18.1	None	2.9	b-group
7	17.9 19.3(0 σ)	None	2.9	b-group b-group
8	20.5(0 σ)	None	2.9	b-overall

Table 6.1 Stages of the refinement of GDH. Initial R-factor = 49.3%, refinement with a 2σ cut off in the reflection file unless stated. Round 5 of the refinement is believed to represent the most crystallographically valid model.

Assessment of model quality

The structure of GDH has been determined to 2.9\AA resolution with an overall R-factor of 19.3% for all reflections between 8 and 2.9\AA . A Luzatti plot (fig. 6.4) (from the X-PLOR version 2.1 manual) indicated a mean positional error of $\approx 0.3\text{\AA}$, which rose sharply beyond $\approx 3.2\text{\AA}$. Figs. 6.5 & 6.6 shows how the refined structure correlated with the initial averaged map and the final $2\text{Fo}-\text{Fc}$ map. The mean RS_FIT (from 'O') score for the averaged map/structure was ≈ 0.6 , whereas the $2\text{Fo}-\text{Fc}$ map gave an improved value of ≈ 0.7 . Within some of the loop regions, the structure did not

correlate very well with the initial averaged map, which may be indicative of disorder within these loop regions, or inadequacies in the mask.

The 3D-1D score (Luthy *et al.*, 1992), which gives a measure of the biochemical sense of the structure, was correlated for the monomer and the tetramer (fig. 6.7). Low scores were obtained for the monomer in region 100-110 (0.1) and 280-300 (0.0).

These regions participate in the subunit/subunit interactions. The score for these

luzatti plot

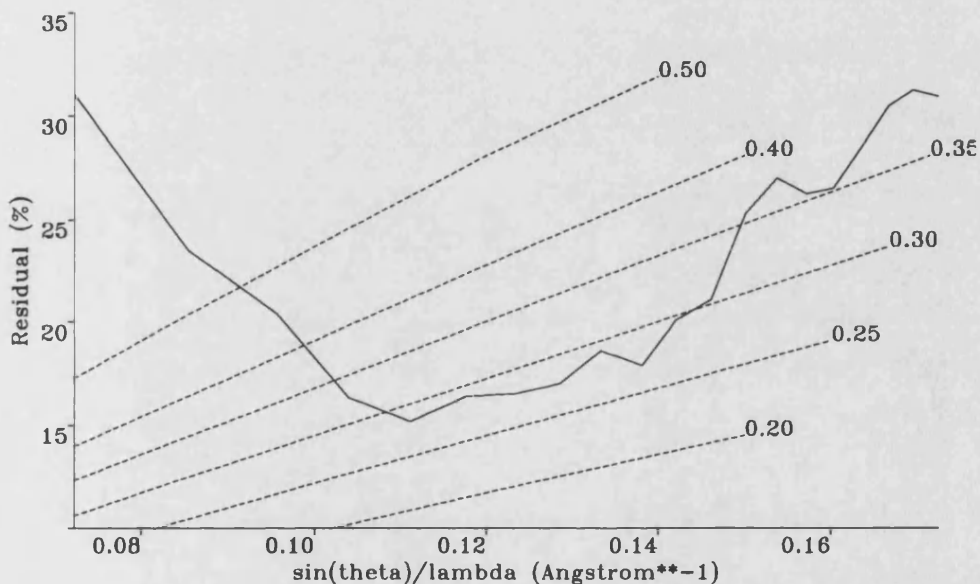


Fig. 6.4 GDH luzatti plot

regions improved when the complete tetramer was taken into consideration.

PROCHECK (Morris *et al.*, 1992) uses various criteria to assess the quality of the structure. A Ramachandran plot displays the residues that exhibit acceptable stereochemistry and highlights the outliers. In the refinement, efforts were made to maximise the stereochemical quality of the monomer only (group11_a) (figs. 6.8 & 6.9). However, as the ncs restraints were eventually removed in the refinement, it is

likely that the other monomers do not exhibit equivalent stereochemistry. This is borne out by the Ramachandran plots of the 3 other monomers (fig. 6.11).

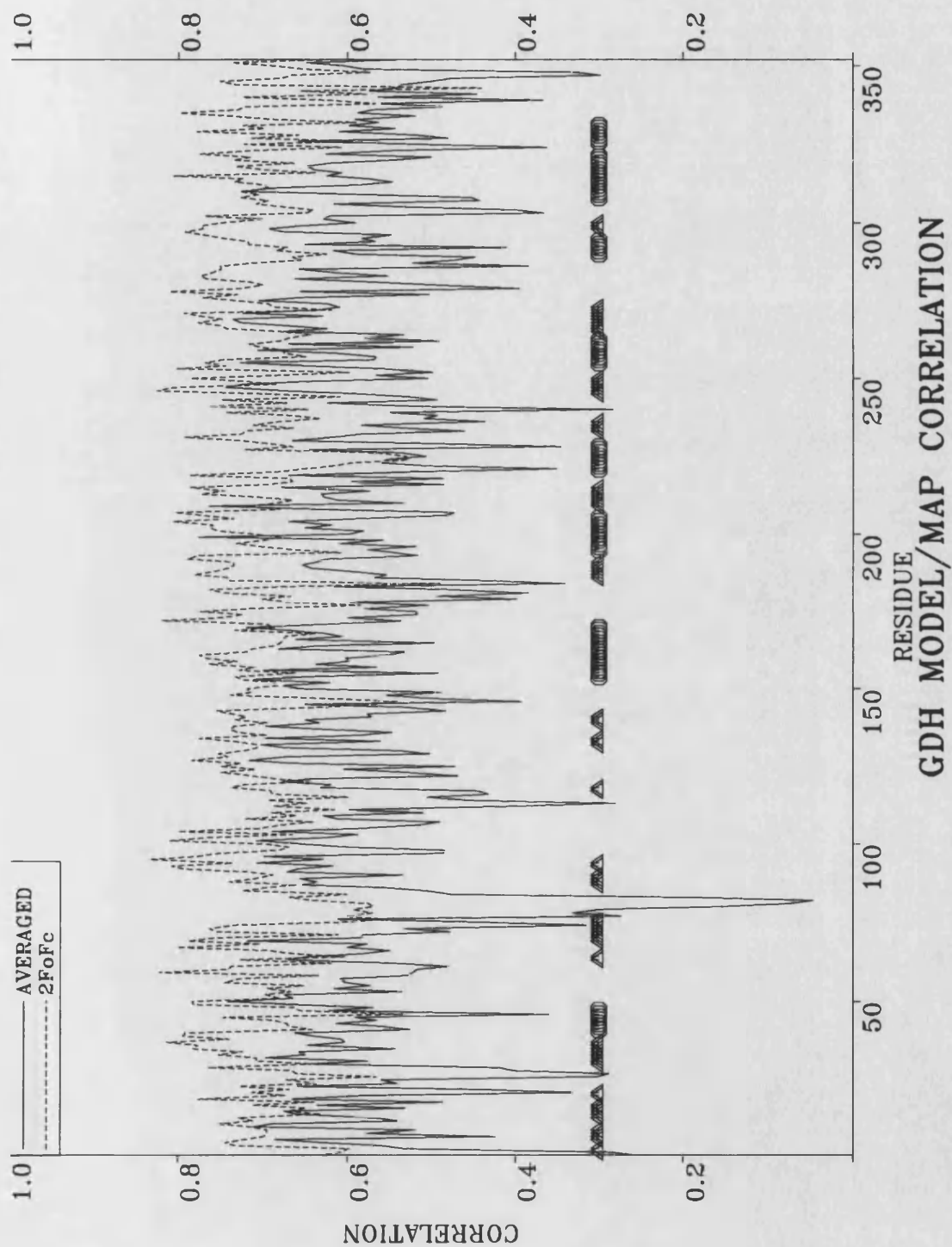


Fig. 6.5 GDH model correlation to the averaged map and 2Fo-Fc map. The secondary structural elements of strands and helices are represented as triangles and circles respectively.

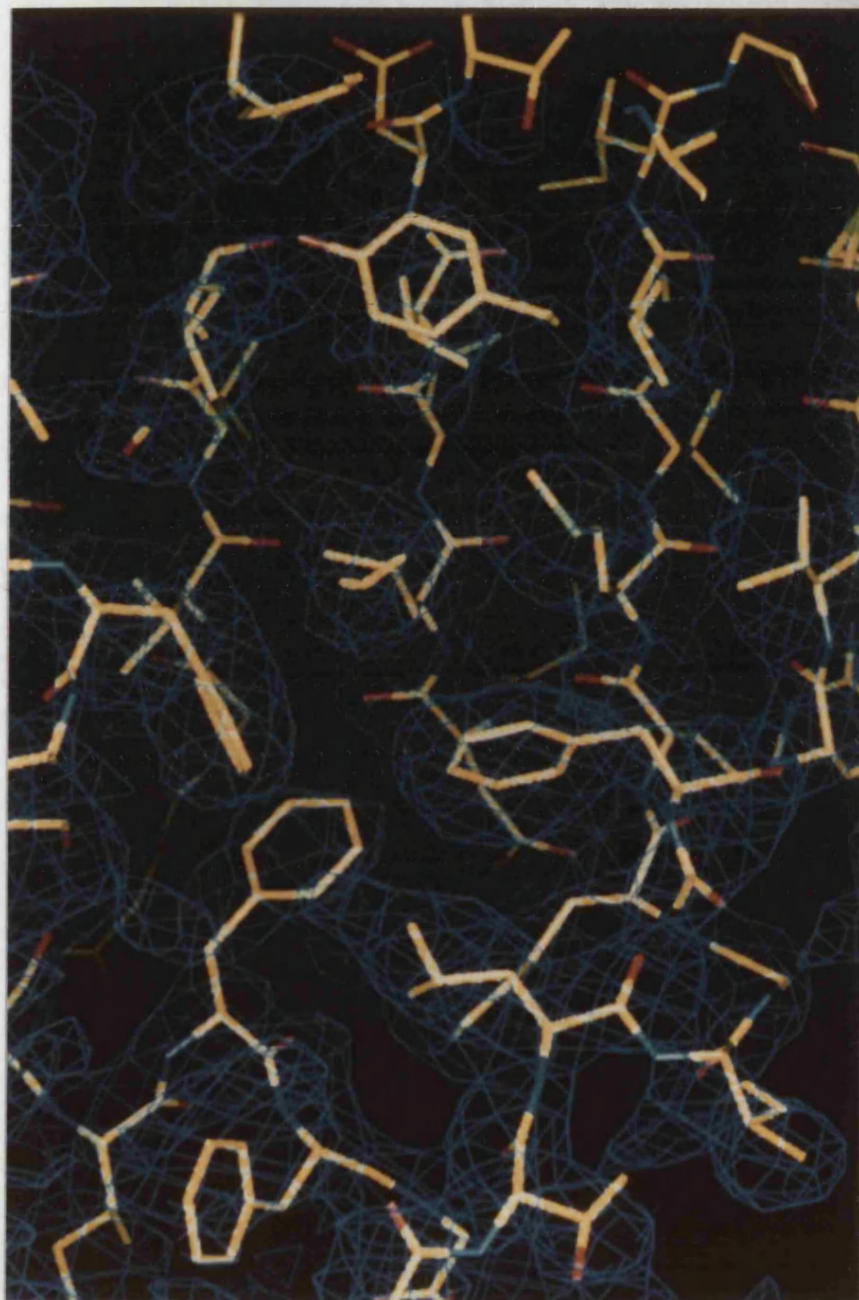


Fig.6.6 A region of the refined structure superposed onto the corresponding $2F_o - F_c$ electron density.

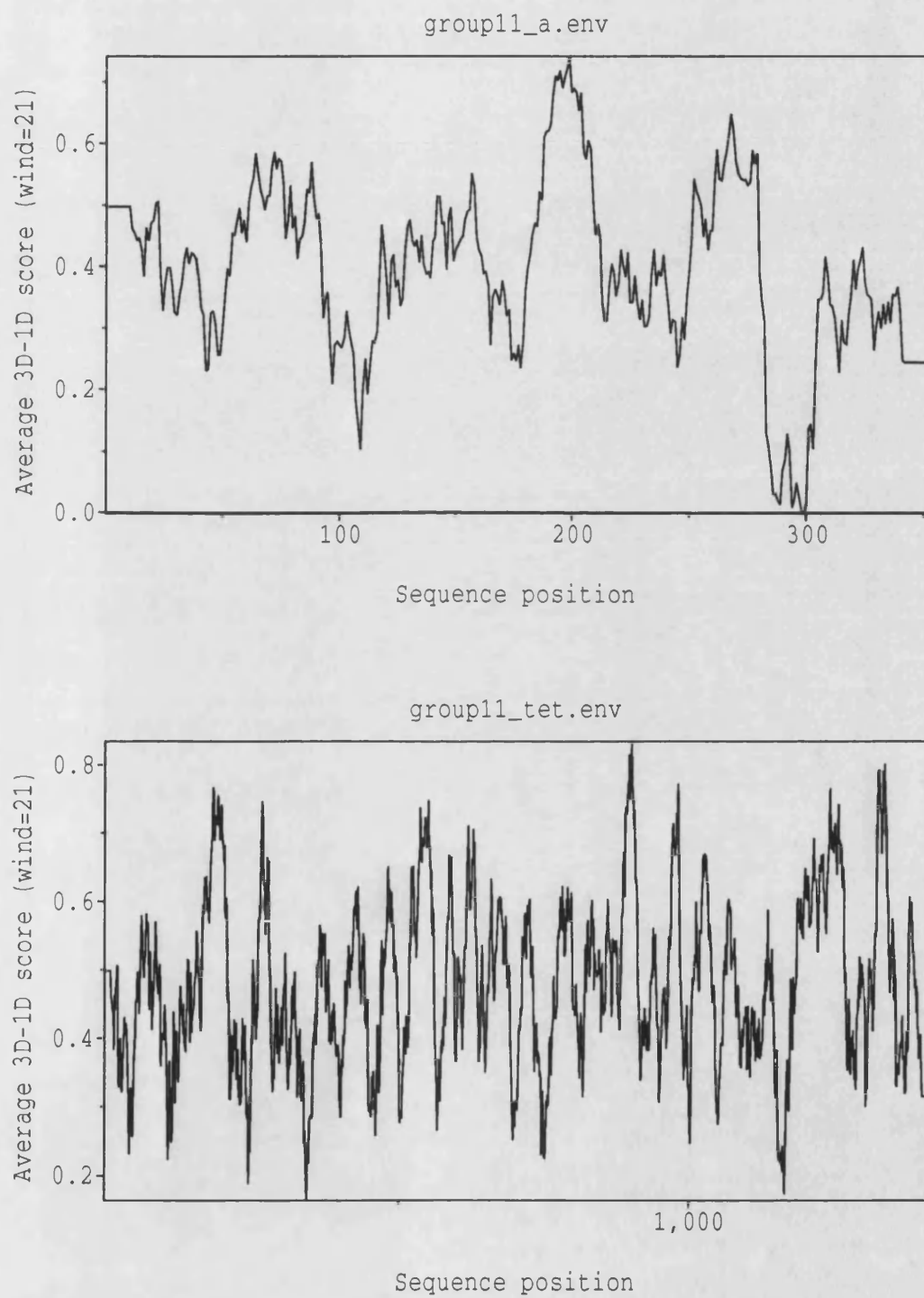


Fig. 6.7 The 3D-1D scores for the monomer and the tetramer.

The Ramachandran plot for monomer A reveals that the majority of the residues are in the most favoured regions and the additionally allowed regions (96.7%). Only one residue, Asp 75, displays unusual stereochemistry - this residue is in a surface loop (75-81). However, it can be seen that many of the residues within this surface loop exhibit precarious stereochemistry. This loop is represented by a discontinuous stretch of density, and exhibits relatively high temperature factors. Furthermore this loop is involved in crystal packing contacts - which may explain why this loop adopts a strained conformation.

Various main chain and side chain parameters (fig 6.9) also indicate that the GDH structure conforms to the required geometrical standards. The rms deviation in the bond length and bond angle from ideality were 0.023Å and 4.576° respectively. The average main chain and side chain temperature factors were 17.61Å² and 17.77Å² respectively. In general, the temperature factor for the main chain and side chain residues were lower for regions within defined structural elements compared to loop regions - as expected (fig. 6.10).

Heavy atom site

The platinum heavy atom compound binds to a methionine 241 in each monomer. Methionine 241 is located on a solvent exposed surface loop between βC and βD within the nucleotide-binding domain (fig. 6.12).

Crystal packing

Examination of the crystal packing (fig. 6.13) revealed that there were no serious clashes between the crystallographically related tetramers, although some contacts were observed (especially loop 75 -81).

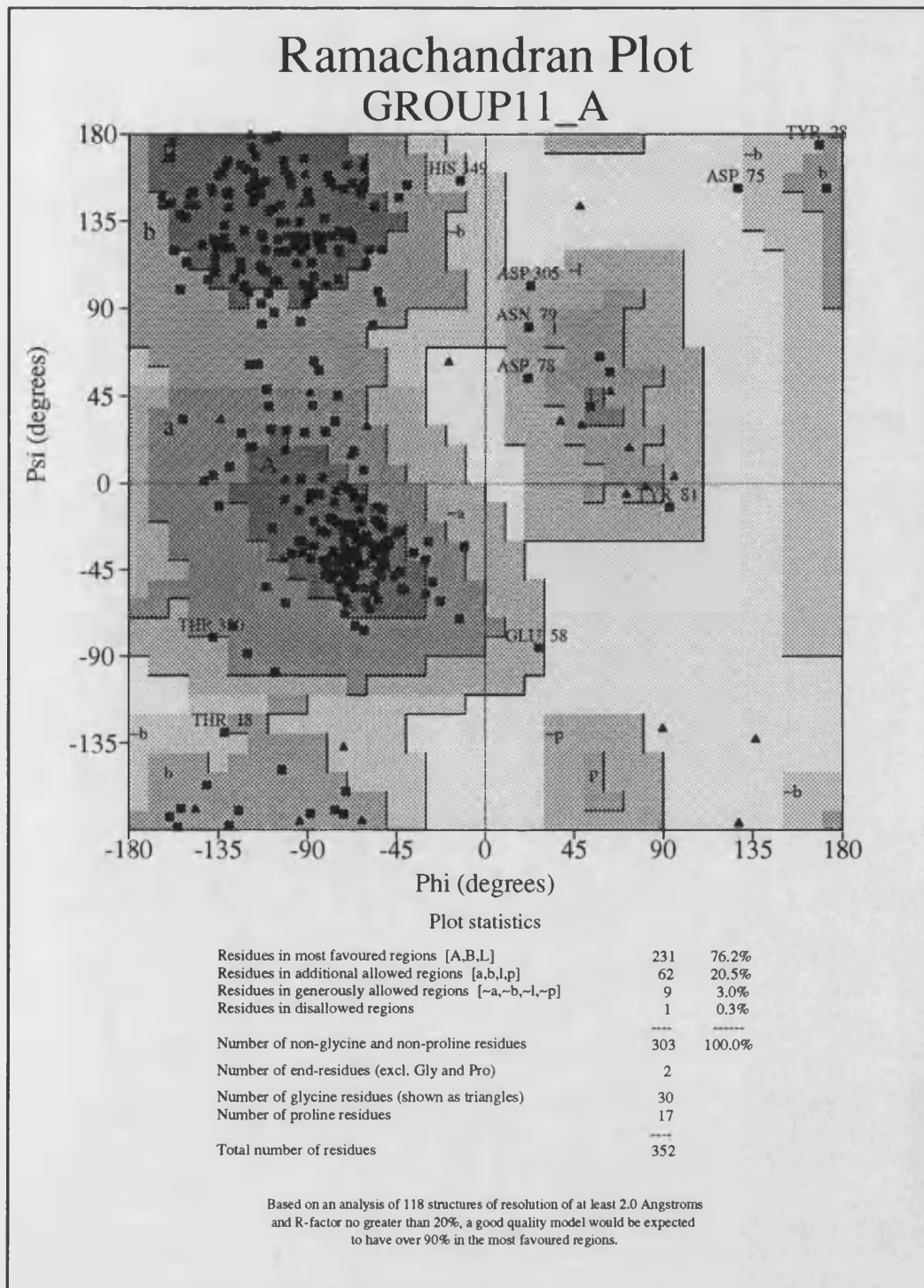


Fig. 6.8 The Ramachandran plot of GDH monomer A.

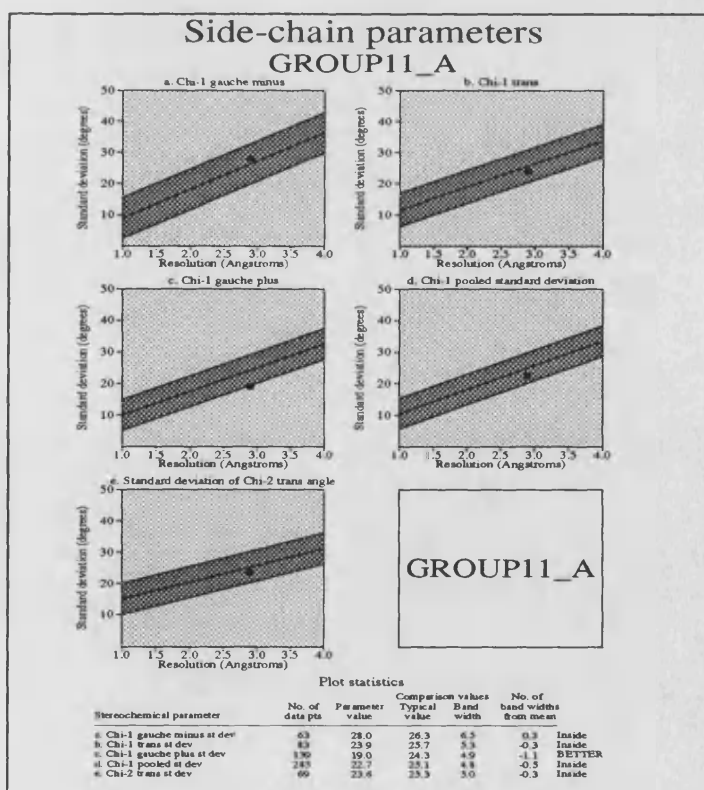
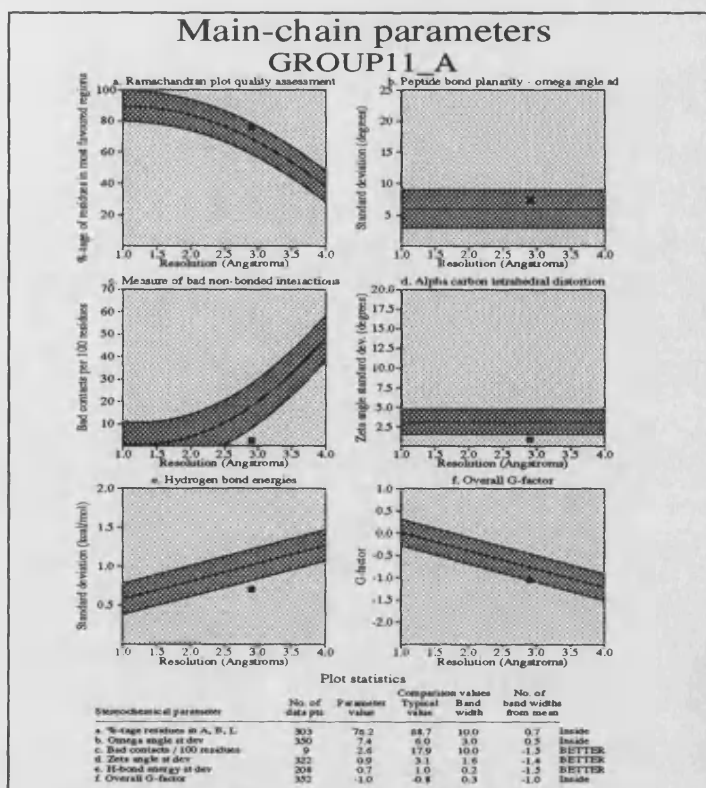


Fig. 6.9 Main chain and side chain parameters of group11_a

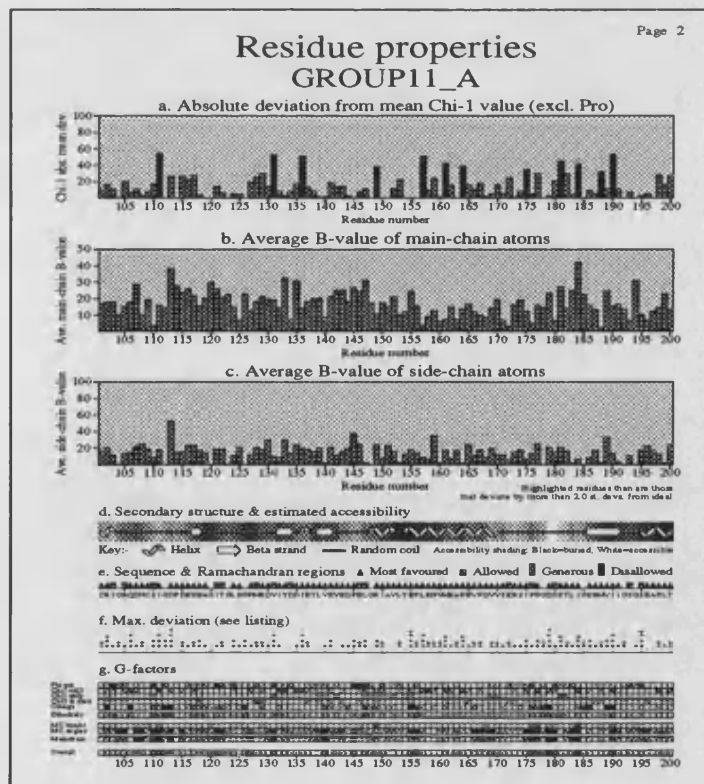
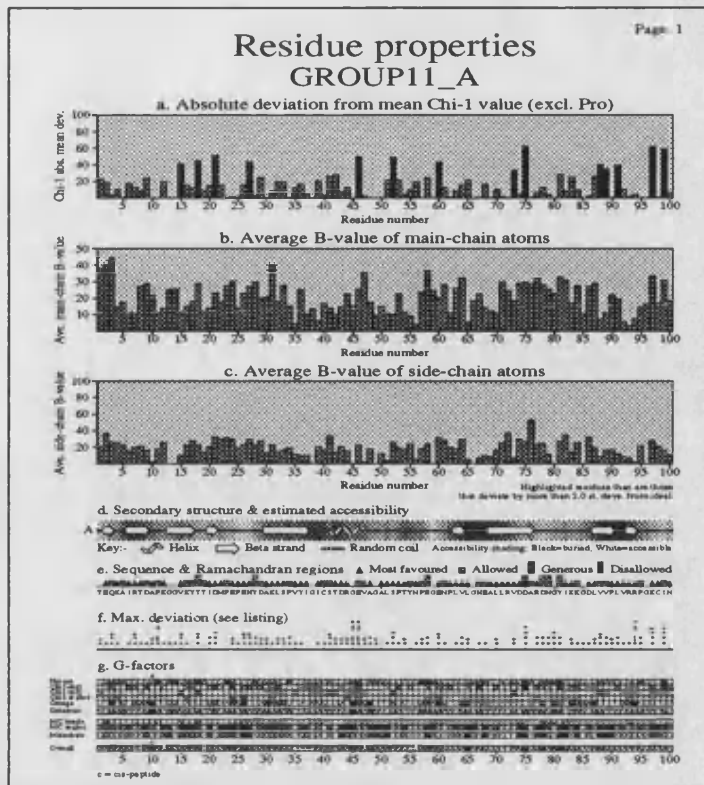


Fig. 6.10 Main chain and side chain temperature factor plot per residue.

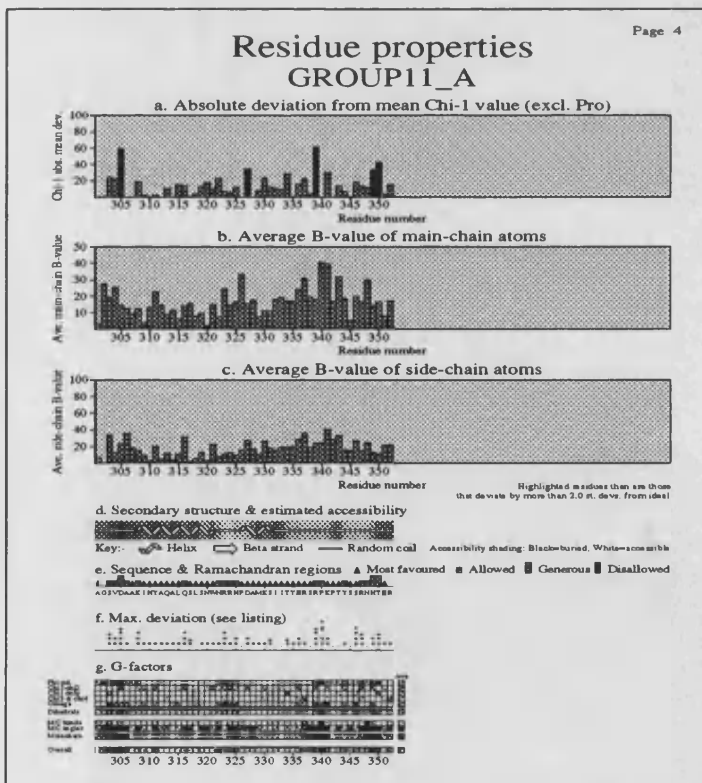
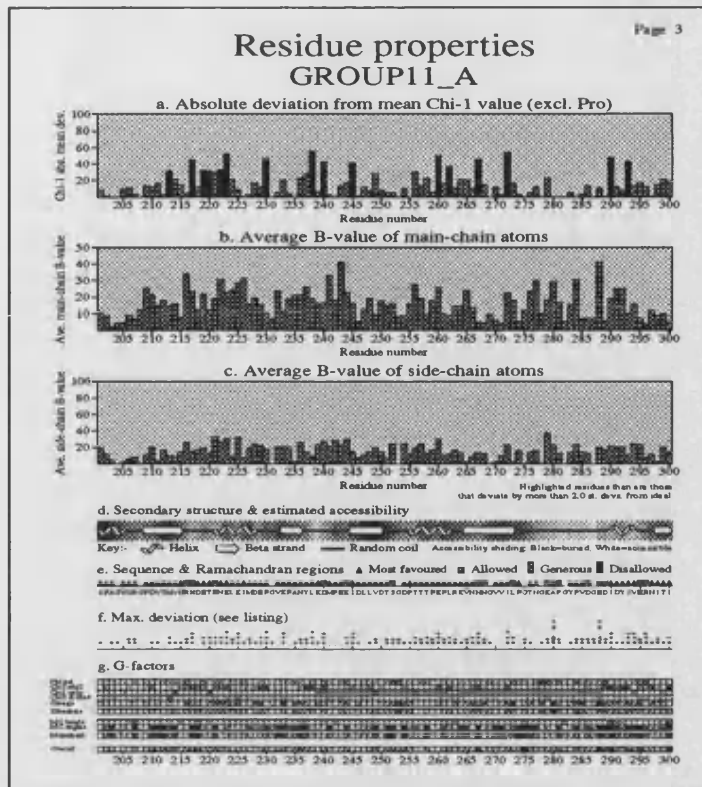


Fig. 6.10 (cont'd) Main chain and side chain temperature factor plot per residue.

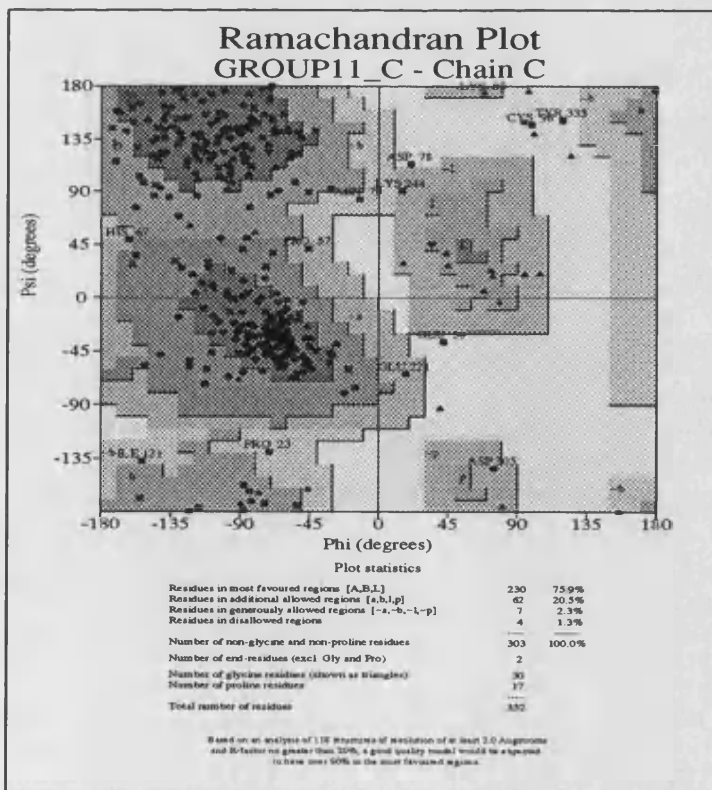
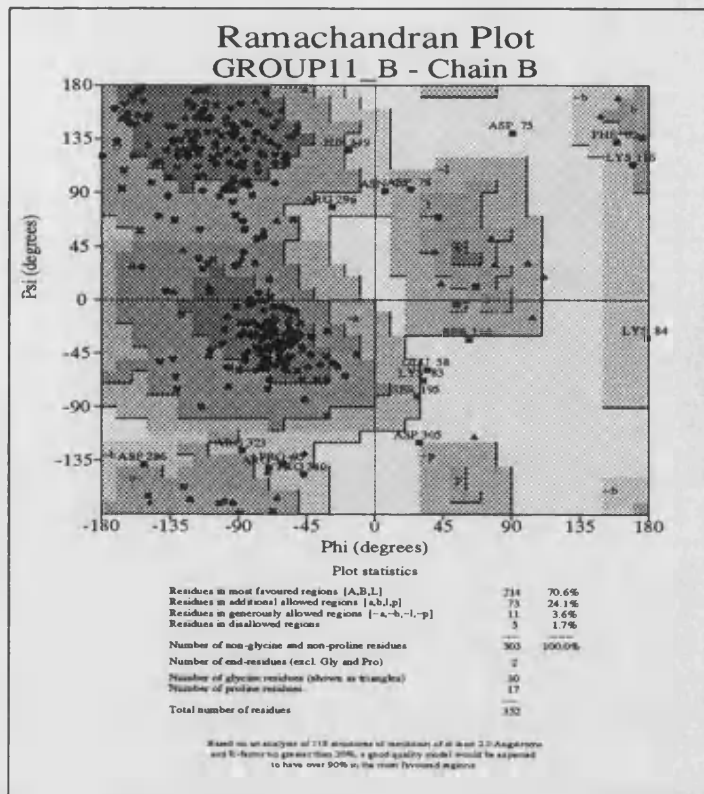


Fig. 6.11 Ramachandran plot of group11_b and group11_c.

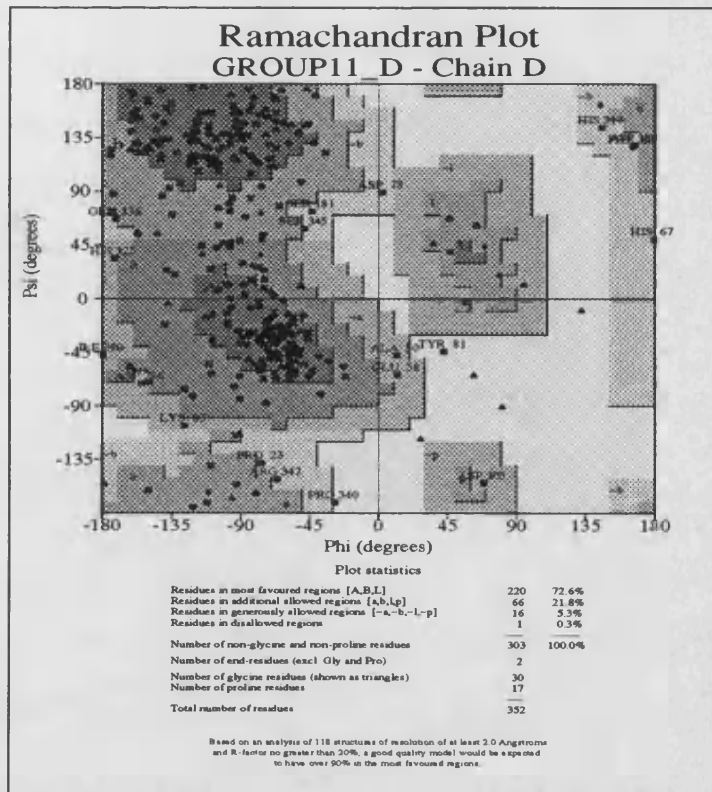


Fig. 6.11(cont'd) Ramachandran plot of group11_d.

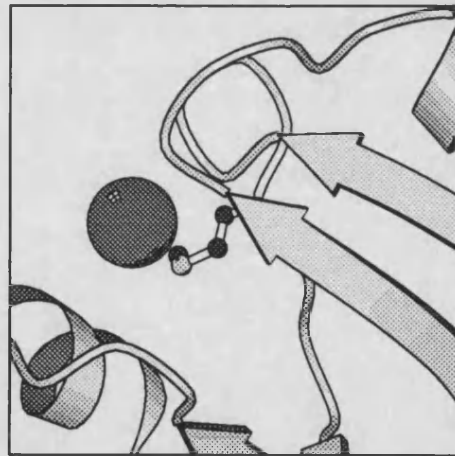


Fig. 6.12 The Platinum heavy atom site ligated to Methionine 241.

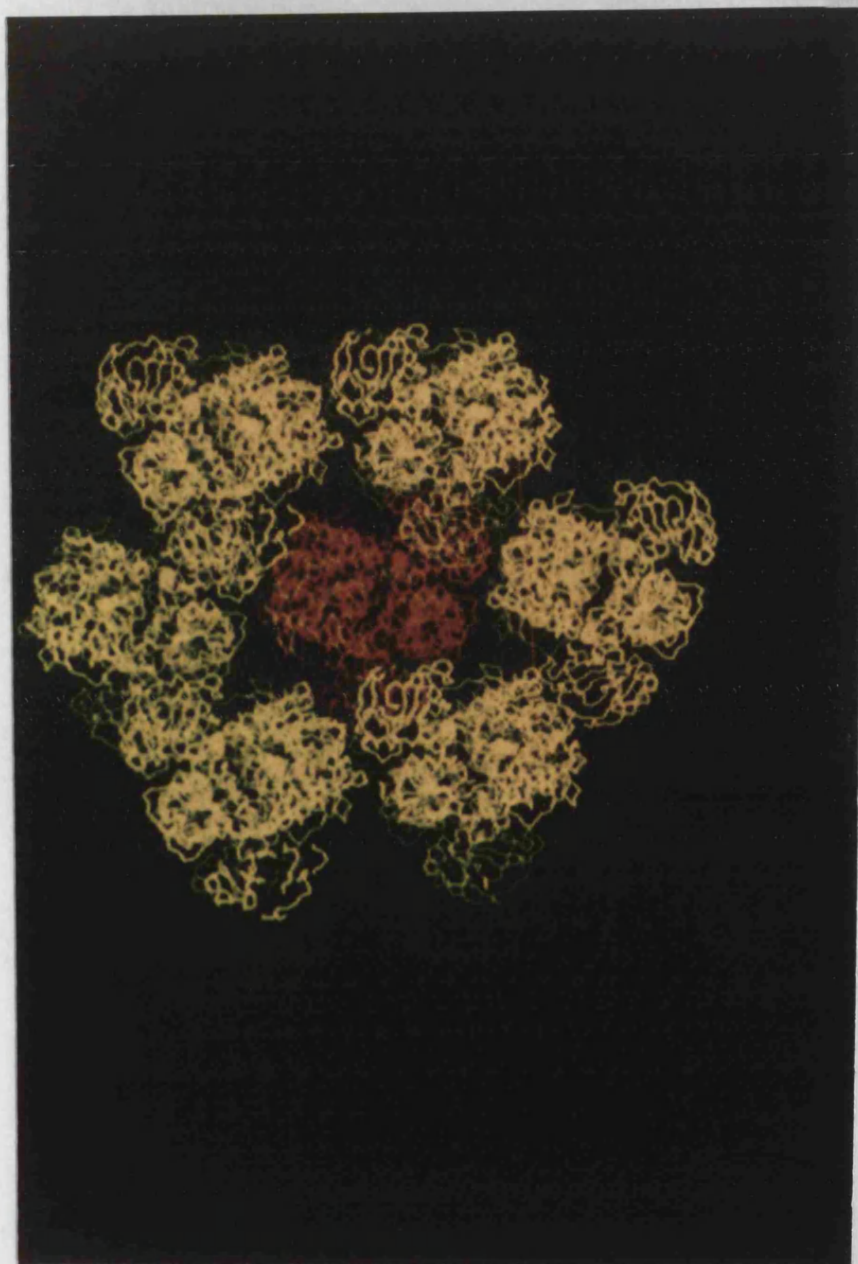


Fig. 6.13. The crystallographically related GDH tetramers.

Chapter 7

THE GDH STRUCTURE

Structural characterisation of the monomer

The GDH monomer consists of two domains, a central nucleotide binding domain (residues 187-300) flanked by the catalytic domains (residues 1-186, 301-352) (fig. 7.1 & 7.2). The monomer has an approximate helical content of 22%, while 25% is β -strand. The catalytic and nucleotide-binding domains are separated by a deep active site cleft that is freely accessible to the solvent. The putative catalytic zinc is at the bottom of the cleft and the lobe containing the putative structural zinc is at the mouth of the cleft.

The nucleotide binding domain

The GDH nucleotide-binding domain is comprised of a β -pleated sheet of six parallel strands (β A to β F) with a marked 100° left handed twist within the sheet, consistent with the nucleotide-binding domain of previously reported structures, and represents the core of the Rossmann fold (Rossmann *et al.*, 1974). This sheet is flanked by 4 helices (α B, α C, α E and α F), which vary in length when compared with the flanking

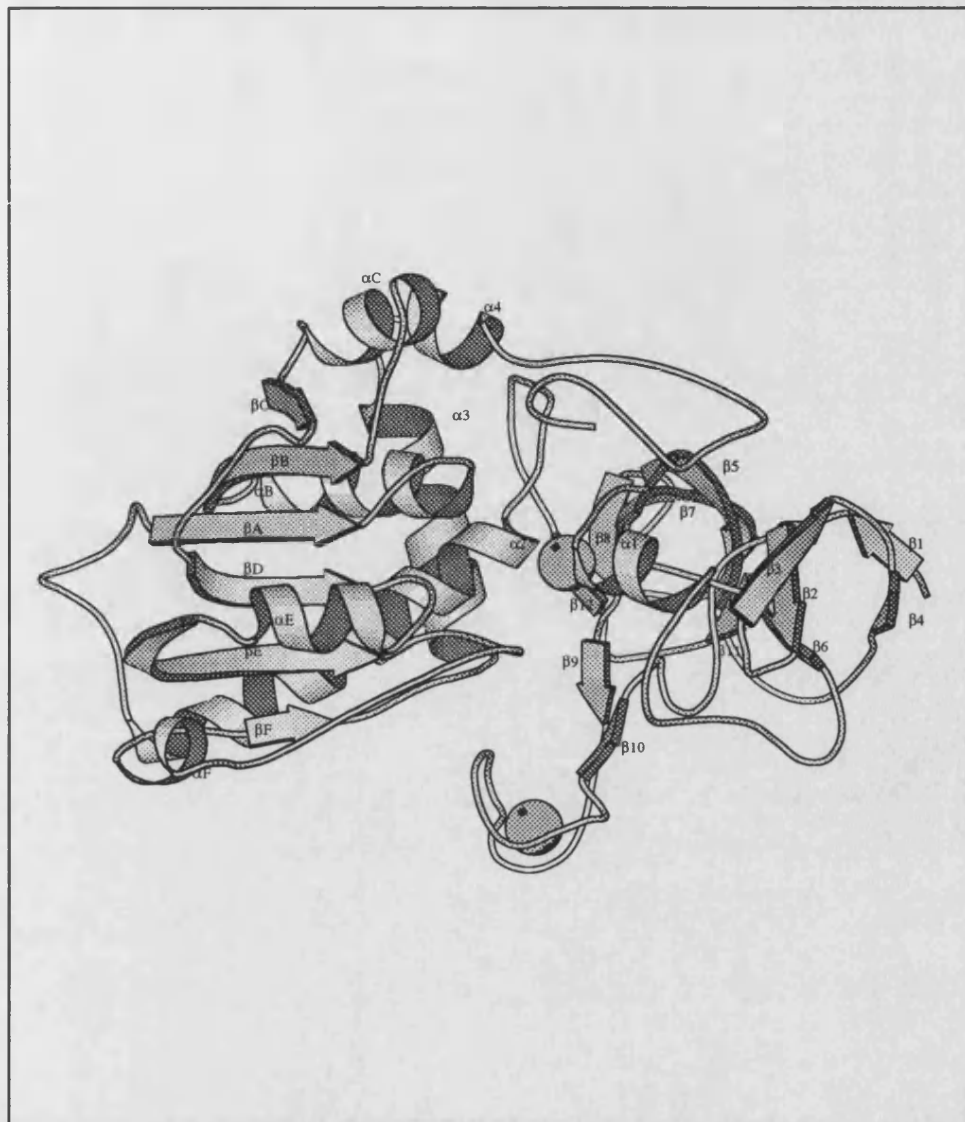


Fig. 7.1 Schematic representation of the GDH monomer drawn using Molscript (Kraulis, 1991) The secondary structure was assigned using DSSP (Kabsch & Sander, 1983). The nomenclature of the GDH nucleotide binding domain was adopted from previously published dehydrogenases. N-terminal catalytic domain: $\beta 1$, 2-4; $\beta 2$, 6-9; $\beta 3$, 13-17; $\beta 4$, 20-22; $\beta 5$, 30-37; $\beta 6$, 62-64; $\beta 7$, 69-76; $\beta 8$, 87-90; $\beta 9$, 92-94; $\beta 10$, 116-118; $\beta 11$, 132-134; $\beta 12$, 139-141; $\alpha 1$, 41-47; $\alpha 2$, 154-174. Nucleotide-binding domain: βA , 186-191; βB , 209-215; βC , 233-236; βD , 245-250; βE , 265-273; βF , 298-300; αB , 195-205; αC , 221-228; αE , 255-262; αF , 290-294. C-terminal catalytic domain: $\alpha 3$, 308-321; $\alpha 4$, 326-331.



Fig. 7.2 Stereo view of the C α tracings of the GDH monomer.

helices of other dehydrogenases, underlying the finding that the β -pleated sheet is the most structurally conserved feature of the nucleotide-binding domain.

The GDH nucleotide binding domain was superposed onto the nucleotide binding domains of a number of dehydrogenases using COMPOSER (Sutcliffe *et al.*, 1987) (table 7.1). Residues within the β A, β B and β C of each dehydrogenase aligned were used to initiate the fitting process (in addition, using only the GxGxxG/A loop as the initial site of structural homology yielded the same result) Fig. 7.3 & 7.4 highlights the conserved structural nature of the nucleotide binding domain.

Structures	length of enzyme (a.a)	nucleotide binding region	length of nucleotide domain (a.a)	Additional structural features
Glyceraldehyde 3-phosphate dehydrogenase 2GPD	334	1→148	148	loop (45-72), loop (76-82), <i>loop (99-104)</i>
Glutathione reductase 3GRS	478	158→293	135	loop & helix (222-233), <i>2strands (247-273)</i>
Lactate dehydrogenase 4LDH	330	22→164	142	loop (68-77), <i>loop & helix (97-119)</i>
Alcohol dehydrogenase 8ADH	374	193→319	126	helix (246-261)
Sorbitol dehydrogenase 1SDG	374	193→319	126	helix (246-261)
Glyceraldehyde 3-phosphate dehydrogenase 2GD1(thermophile)	333	1→148	148	loop (45-72), loop (76-82), <i>loop (99-104)</i>
6-Phosphogluconate dehydrogenase 1PGD	469	3→128	125	longer helix (34-48), <i>small helix and loop (75-79)</i>
Glutamate dehydrogenase	449	231→368	137	helix and loop (290-305)
GDH	352	186→300	114	n/a

Table 7.1 The nucleotide binding domains of a number of dehydrogenases: The beginning of β A

and end of β F were used to define the nucleotide binding domain. The length of the domains and

their additional regions are displayed. Additional regions in italics refer to regions beyond β D.

Abbreviations represent the Brookhaven codes for the respective enzymes.

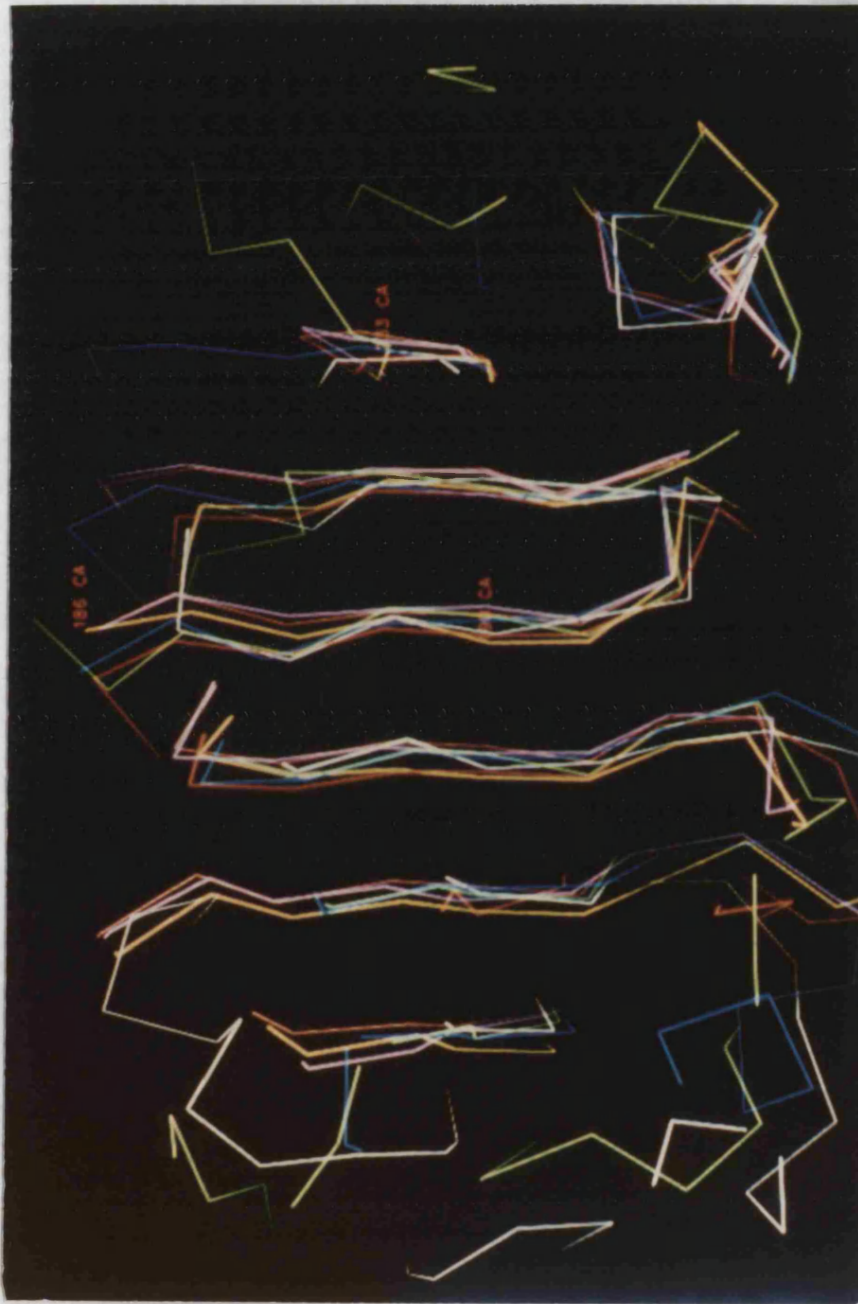


Fig. 7.3 The nucleotide binding domain superposition - showing the markedly conserved nature of the β -sheet.

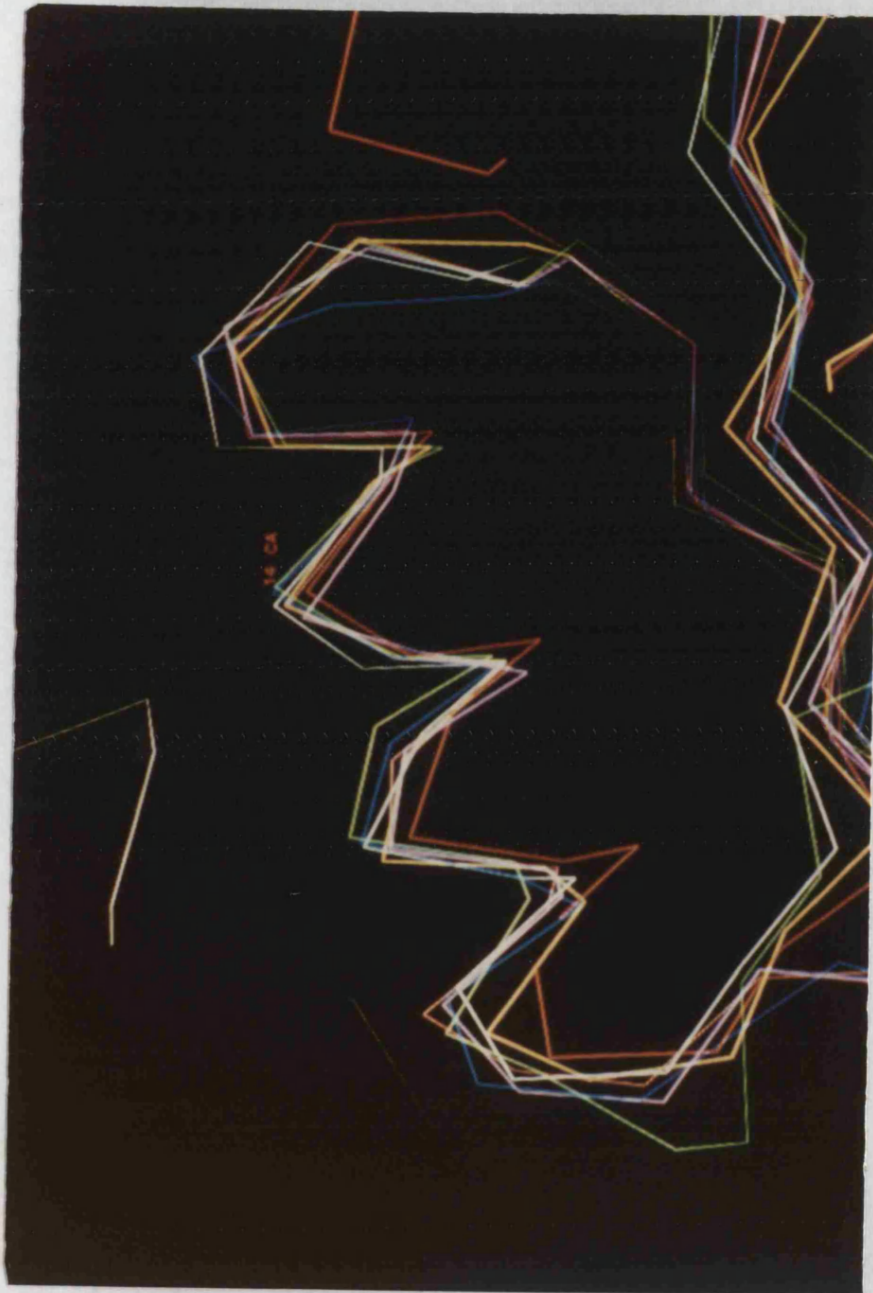
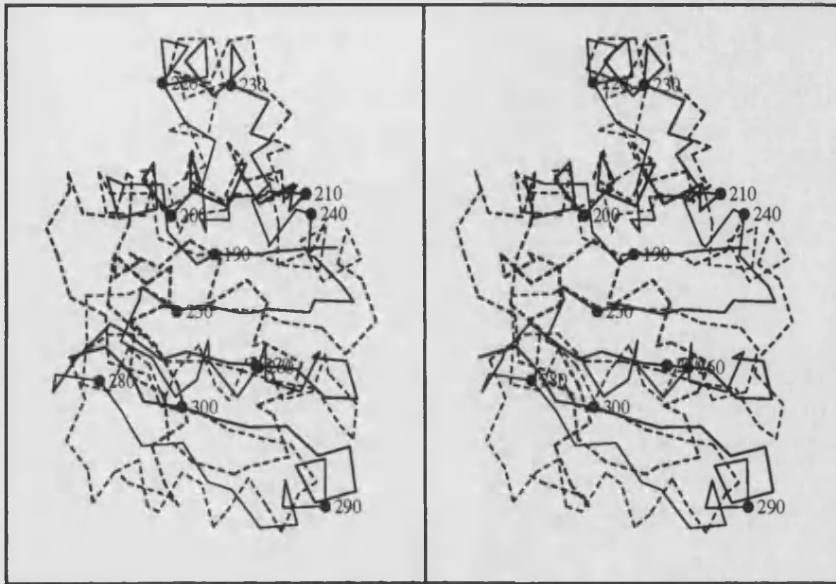
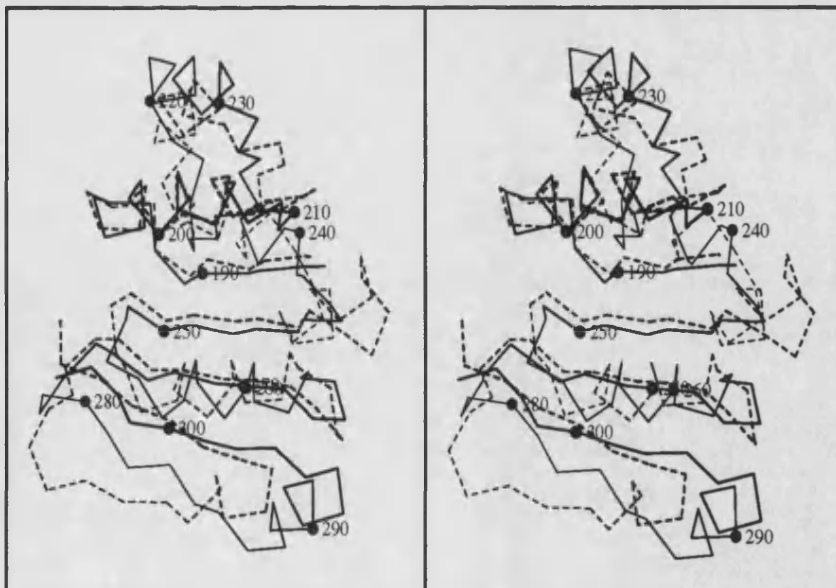


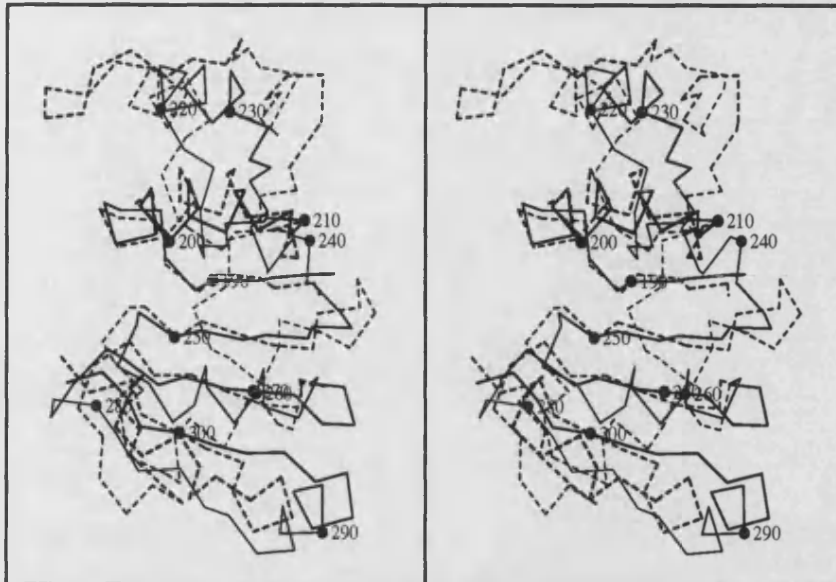
Fig. 7.4. The nucleotide-binding domain superposition - the structurally conserved α B helix.



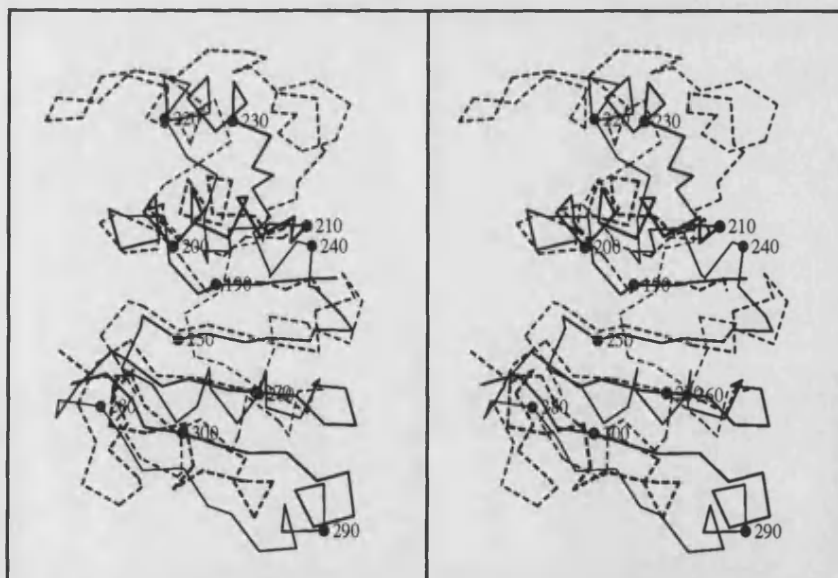
a) 1PGD & GDH



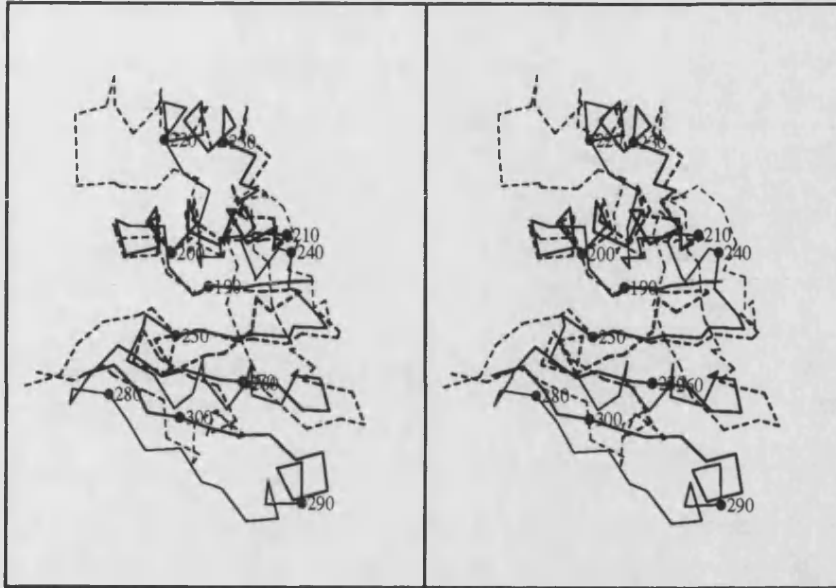
b) 1SDG & GDH



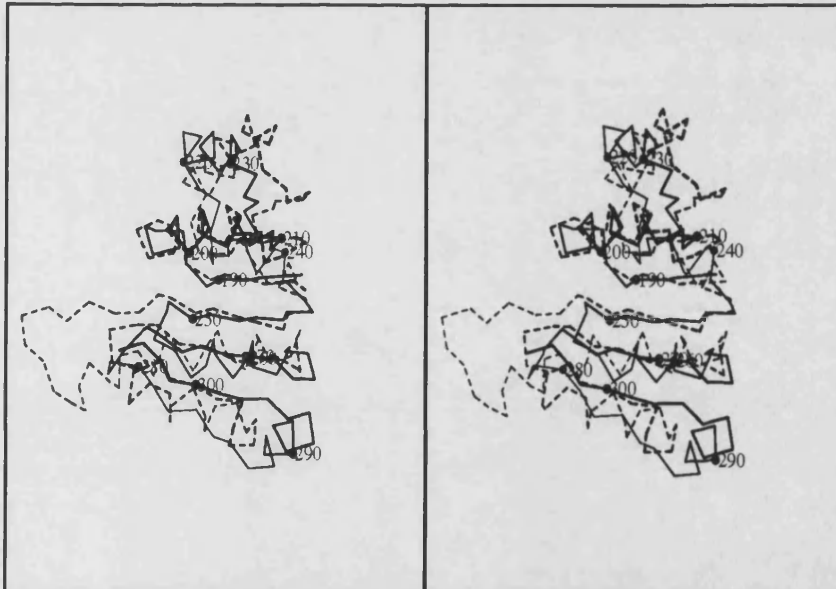
c) 2GD1 & GDH



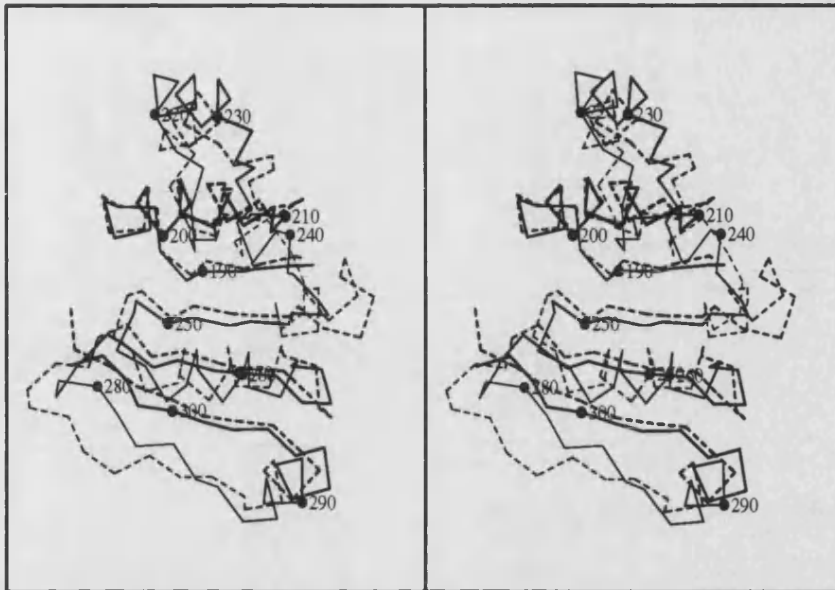
d) 2GPD & GDH



e) 3GRS & GDH



f) 4LDH & GDH



g) 8ADH & GDH

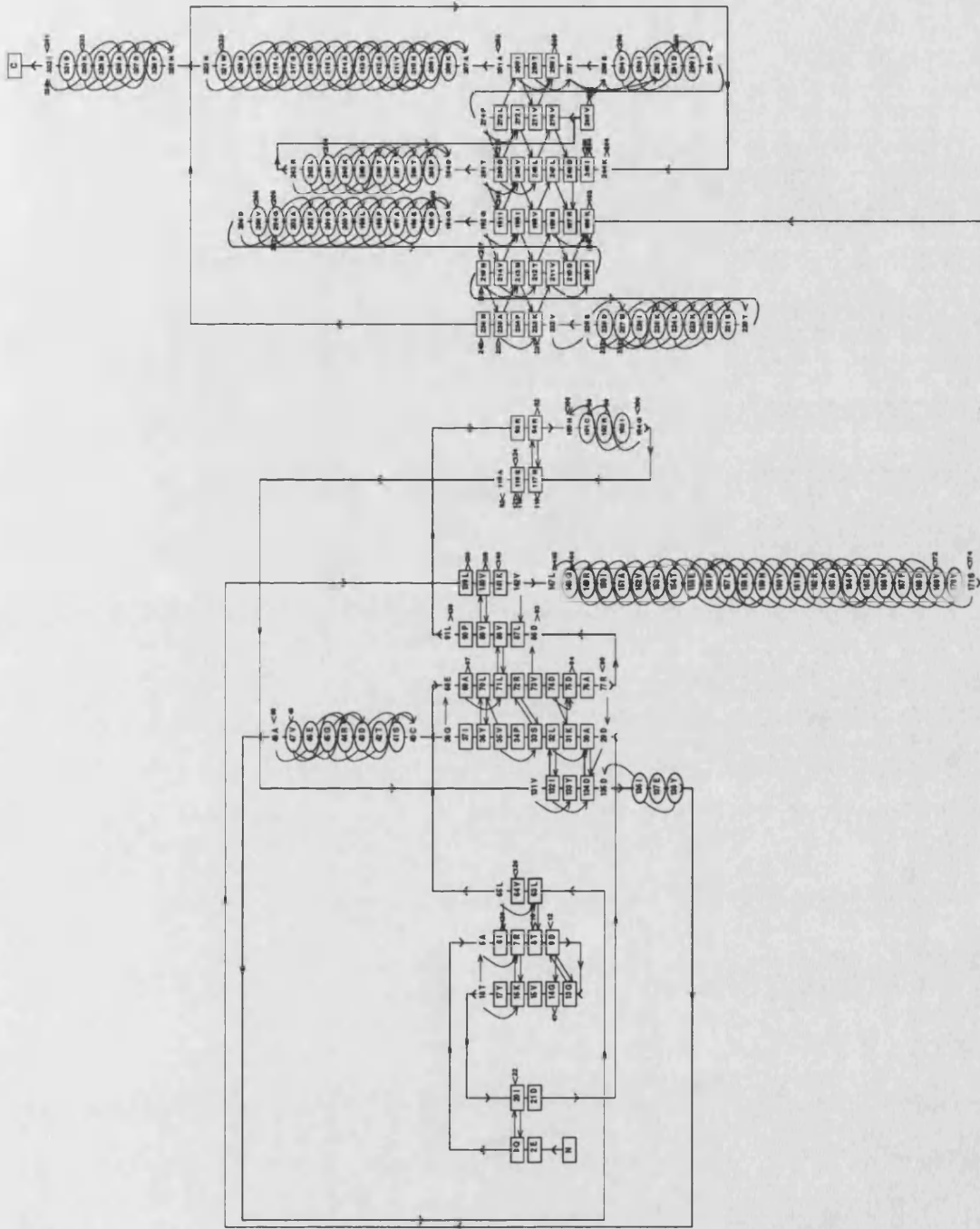
Fig. 7.5 Gallery of stereoplots of the pairwise nucleotide binding domain superpositions. In each plot, the GDH is numbered and represented in solid black lines. The dashed lines in plots a) to g) are 1PGD, PDB1SDG, PDB2GD1, 2GPD, 3GRS, 4LDH, 8ADH respectively.

Putative minimal functional unit

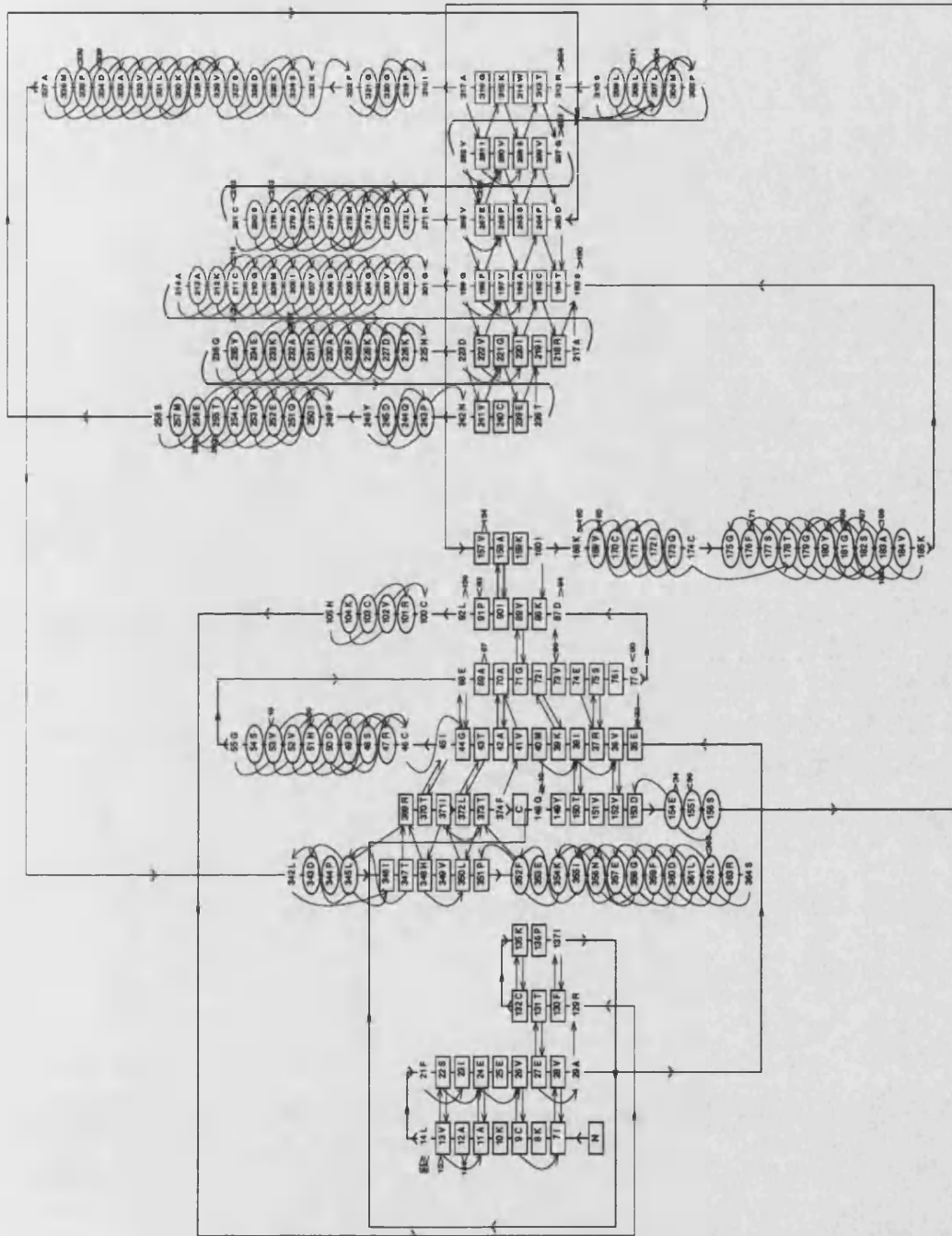
The function of the nucleotide binding domain is to bind the cofactor. Regions that were structurally homologous within this dehydrogenase superposition were used to define the *minimal functional unit* (Russell & John, 1994) of the nucleotide binding domain (excluding glutamate dehydrogenase - it possesses a reverse orientation of the β C strand); that is the minimal amount of the secondary structural elements required to bind the cofactor. The superposition revealed that β A \rightarrow β D, and α B were the conserved features of the nucleotide binding domain. Any structural elements between β A and β D that did not represent the *minimal functional unit* were deemed as *unnecessary* when concerning the binding of the cofactor. In terms of amino acid length, GDH has the smallest nucleotide binding domain. Inspection of structural superposition revealed that the nucleotide binding domains listed in table 7.1 contained additional secondary structural elements that are not present in the GDH nucleotide binding domain. The significance of this observation is discussed in Chapter 9 (Putative Archaeal features). The stereoplots in Fig. 7.5 clearly show the additional elements present in the nucleotide binding domains (with respect to *the minimal functional unit*).

The catalytic domain

Unexpectedly, structural analysis revealed that the catalytic domain of GDH shared extensive homology to the catalytic domain of LADH (Eklund *et al.*, 1976). A comparison of the overall topologies (HERA output, fig. 7.6a,b), the superposed C α tracings of GDH and



a) GDH



b) LADH Fig.7.6 Secondary structural elements were defined using DSSP. Topology diagram produced using HERA, ellipses and boxes represent helical regions and β -strand regions respectively.

a) The topology of GDH. **b)** The topology of LADH. The Rossmann fold is positioned at the top of the page.

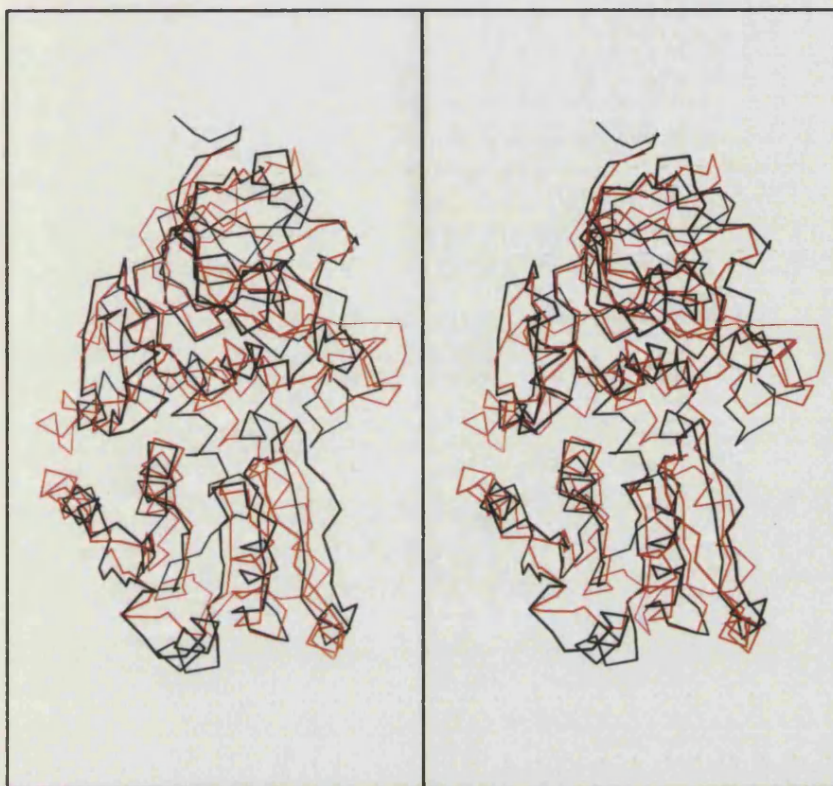


Fig. 7.7 Stereoview of superposed α traces of GDH (red) and ADH (black).

LADH monomers (fig. 7.7), and a structurally based sequence alignment (Fig. 7.8) demonstrates the structural homology between the enzymes. The GDH catalytic domain is described by comparing it to that of LADH. In addition, structural similarities and differences between GDH and LADH may serve to strengthen or refute the monomeric models of yeast ADH (YADH) (Jornvall *et al.*, 1978) and sorbitol dehydrogenase (SDH) (Eklund *et al.*, 1985), which are tetrameric, and based on the LADH structure.

Two antiparallel β -pleated sheets, β I and β II represent the major secondary structural elements of the GDH catalytic domain. The sheet β I, which is equivalent to sheet β III



Fig. 7.8 Structurally based sequence alignment of GDH and LADH. The secondary structural elements of LADH (Eklund and DSSP defined respectively) and GDH are listed next to their respective sequences. Conserved residues are darkly shaded, conservative changes are lightly shaded. Ligands to the zinc are in bold. Boxed regions represent structurally equivalent regions.

of LADH, is comprised of the strands $\beta 3$, $\beta 2$ and $\beta 6$, while the sheet βII , which is equivalent to sheet βII of LADH, is comprised of strands $\beta 5$, $\beta 7$, $\beta 8$, $\beta 11$ and $\beta 12$. There is no DSSP defined secondary structure beyond helix 4 of the GDH catalytic domain, which is the reason for GDH lacking the third catalytic sheet present in LADH (see HERA output). The cylindrical nature of sheets βII & βIII of the LADH structure are also evident in the equivalent sheets of the GDH structure.

The main area of marked structural difference between GDH and LADH is 25 amino acids (115-140), which contains a 21 amino acid surface loop in LADH. This surface loop of LADH is C-terminal to the structural zinc lobe. It is significant that the main proposed region of difference between LADH and YADH and SDH models is also this surface loop (a 21 amino acid gap in the alignment). This is therefore one feature of the GDH structure which validates the YADH and SDH models.

The remaining structural differences between GDH and LADH monomers are found at the loops connecting the major secondary structural elements.

The Zinc sites

A striking feature of the catalytic domain is the presence of the 2 zinc ions, a putative catalytic and structural zinc. The location of zinc in GDH was determined (and see chapter 10) by contouring the difference Fourier electron density map at a high standard deviation (10σ), resulting in only two areas of electron density within the

subunit. These areas of density corresponded to the zinc sites of the structural and catalytic zincs within the LADH structure.

The catalytic zinc

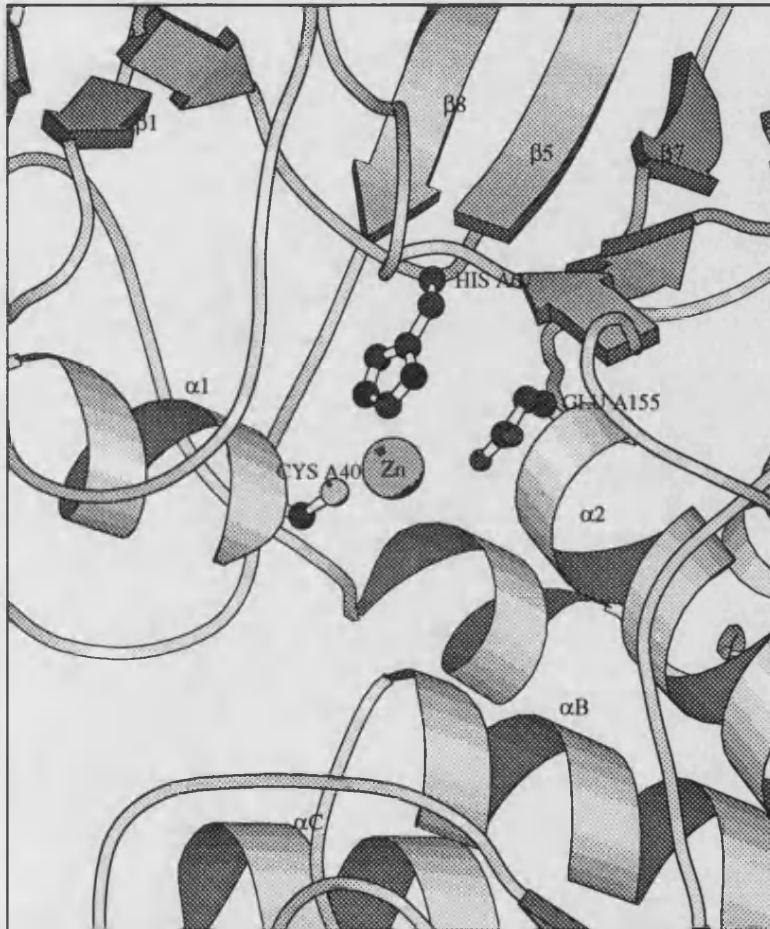


Fig. 7.9 Schematic representation of the ligands to the catalytic zinc in GDH.

The ligands to the catalytic zinc, Cys 46, His 67 and Cys 174, are highly conserved throughout the Zn-ADH family. Two different sequence alignments between SDH and LADH have been proposed: the first (Jornvall *et al.*, 1984) conserves Cys 174 of SDH

as the ligand to the catalytic zinc, resulting in gaps in the alignment between residues 115-190. The alignment according to Eklund *et al.* (1985), which was based on the LADH structure, does not conserve this cysteine as a ligand and rejects the Jornvall sequence alignment as being incompatible with the function of SDH. Instead, Glu 174 was predicted to be the ligand to the catalytic zinc. The Cys 174 ligand of LADH is found in the interconnecting helix (α_2), whereas in GDH the equivalent helix does not contain a cysteine residue available for zinc coordination. Instead GDH was found to contain a glutamate (Glu 155) ligating to the zinc instead of the Cys 174 of LADH (fig. 7.9). The Glu 155 of GDH occurs in a structurally equivalent position to the Cys 174 of LADH and indeed to the Glu 174 of the SDH model, representing another feature of the GDH structure that validates the SDH model.

The structural zinc

Not all Zn-ADHs possess the structural zinc, and its role is largely undefined. Within the structural zinc containing enzymes, the structural zinc ligands (Cys 97, Cys 100, Cys 103 and Cys 111 (LADH numbering)) are well conserved. The putative structural zinc of GDH is contained within a lobe similar to that of LADH (fig. 7.10). However the orientation of this structural lobe with respect to the subunit is markedly different to that of LADH, which can be explained by the difference in the fold of the enzymes beyond this lobe. Furthermore, the seven residue loop between the 2 cysteines in LADH (residues 104-110) obscures the active site cleft, whereas in GDH this corresponding loop (residues 102-108) protrudes into the active site cleft. Consequently, the lobe in GDH is less

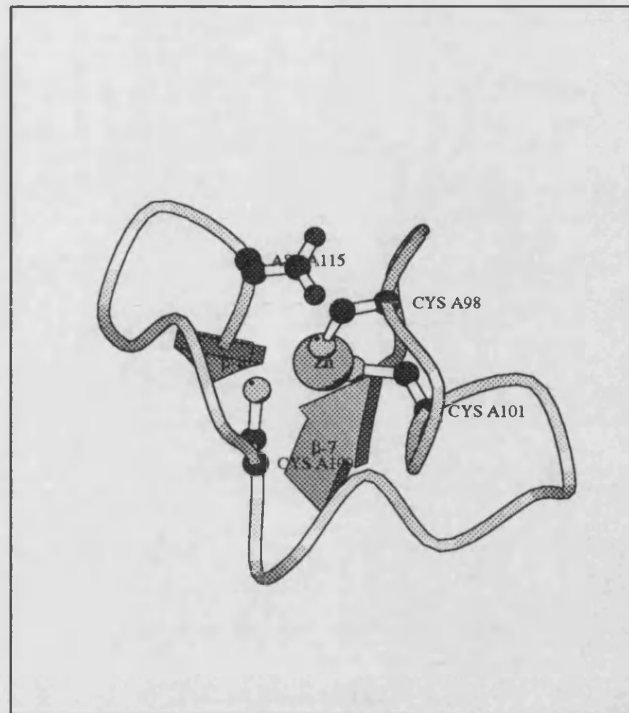


Fig. 7.10 Schematic representation of the ligands to the structural zinc in GDH.

exposed to the solvent. In LADH the structural zinc is ligated to four cysteines, however GDH contains only three cysteines in this lobe (Cys 98, Cys 101, Cys 109). Structural analysis of GDH reveals that the fourth ligand is Asp 115 (fig. 7.10). Interestingly a Zn-ADH from the thermophilic archaeon *Sulfolobus solfataricus* (SSADH) (Ammendola *et al.*, 1992) also contains three cysteines in this region. From the sequence alignment (fig. 7.11), the obvious candidate for the fourth ligand is Glu 98 in SSADH. However in this sequence alignment, the corresponding residue in GDH is a proline, which is unable to ligate to the zinc. In the sequence alignment of SSADH and LADH, there is a notable surface loop gap C-terminal to the structural lobe. This may permit an orientation of the SSADH structural lobe similar to that

298-304 and 311-316 in LADH, appear to be the main region of subunit-subunit interaction.

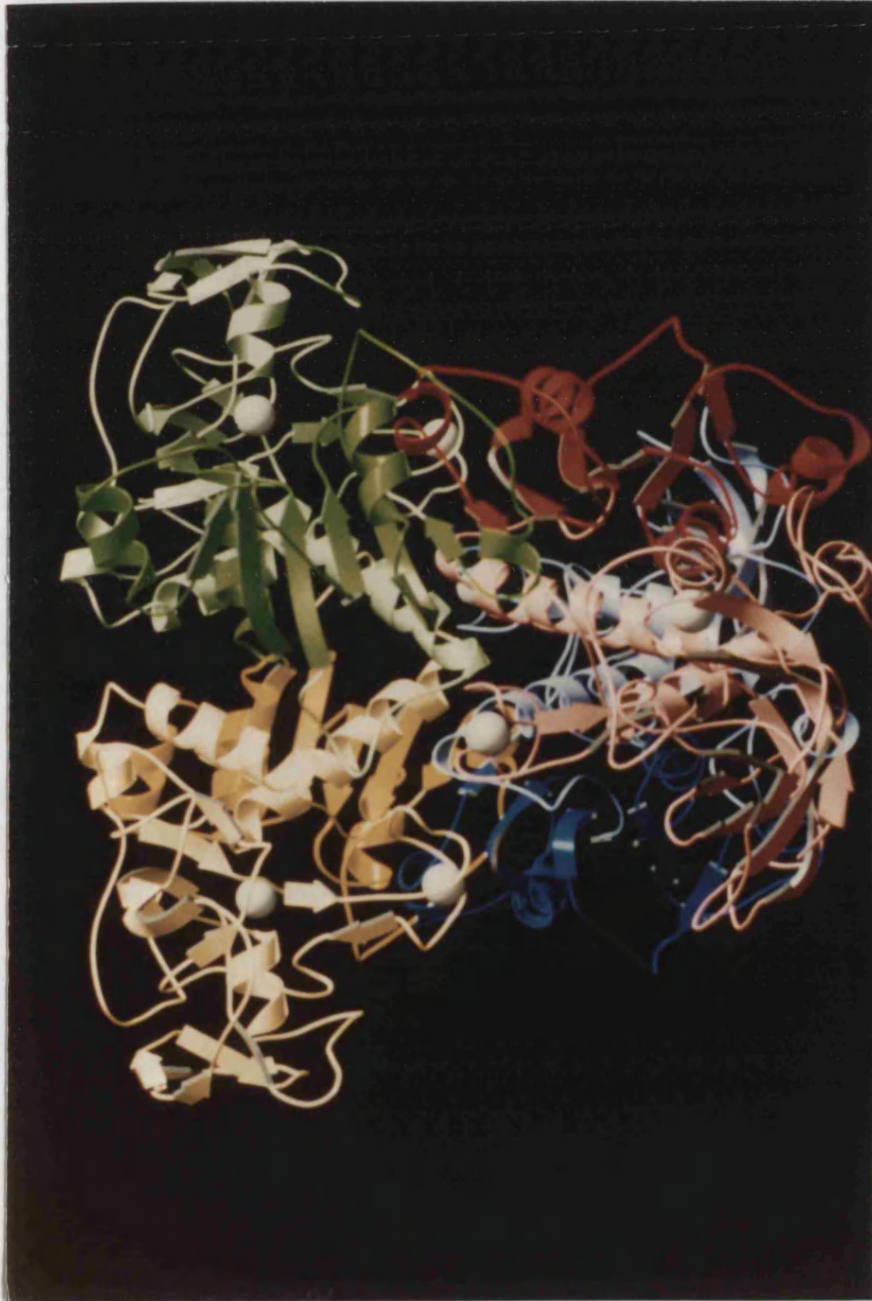


Fig. 7.12a

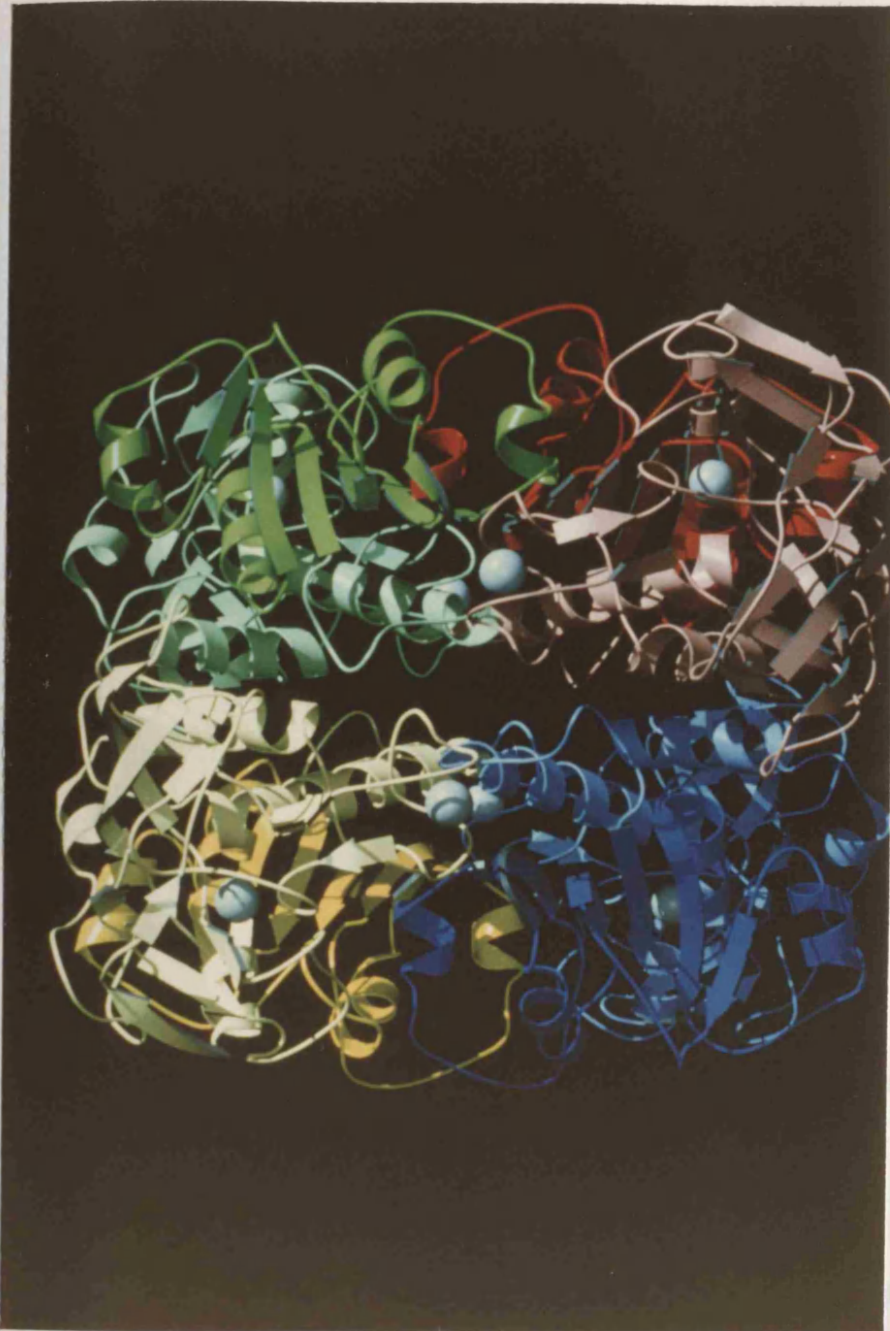


Fig. 7.12b) Schematic representations of the GDH tetramer related by a 45° rotation of the vertical axis (Drawn with MOLSCRIPT and Raster 3D (E Merritt, unpublished program)). Each monomer is shown in a different colour, with the nucleotide-binding domain in a darker shade than the rest of the monomer. Zinc ions are shown as spheres.

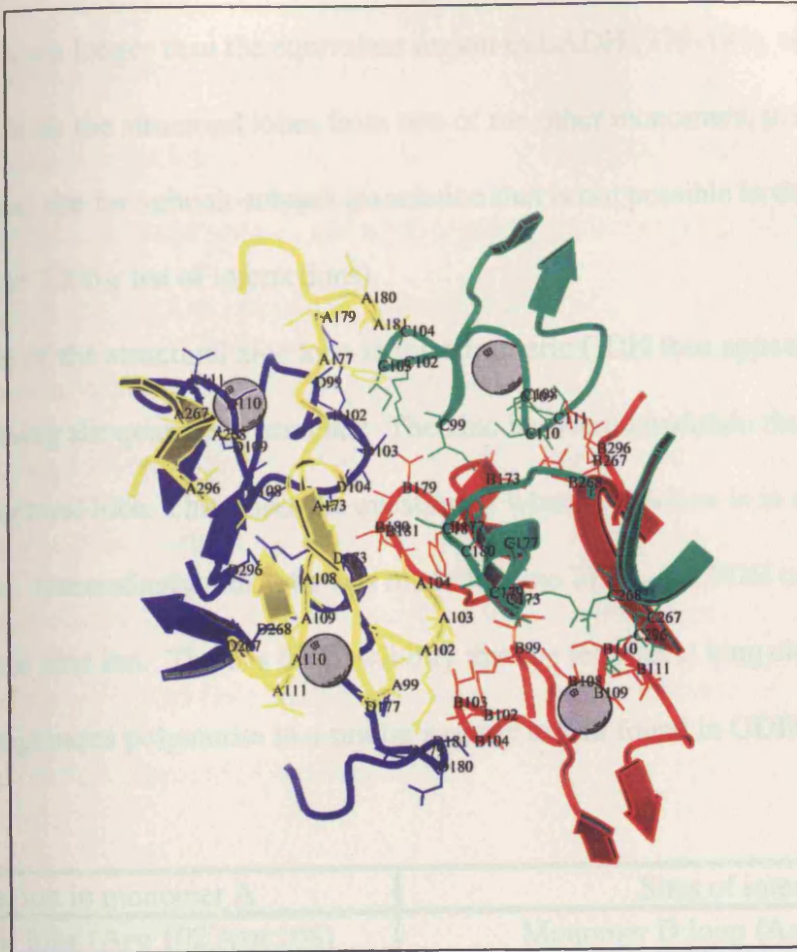


Fig. 7.13 Subunit interaction via the structural lobe. The monomers are coloured coded as in the tetrameric GDH figure.

There is a marked increase in the hydrophobic nature of this nucleotide binding domain interaction in GDH in comparison to LADH, which is a putative thermophilic feature of GDH (see chapter 9).

The zinc-containing structural lobe participates in the other site of subunit-subunit interaction in the tetrameric GDH (fig. 7.13). In GDH, the interconnecting helix ($\alpha 2$) is 2.5 turns longer (nine residues) than the equivalent helix in LADH. This

interconnecting helix of GDH (154-174) and the following carboxy-terminal loop, both of which are longer than the equivalent region in LADH (175-184), enable the loop to interact with the structural lobes from two of the other monomers, providing an additional site for subunit-subunit association that is not possible in the LADH dimer (see table 7.2 for list of interactions).

The role of the structural zinc lobe in the tetrameric GDH thus appears to be that of maintaining the quaternary structure. The zinc ion would maintain the conformation of this structural lobe. This raises the question of what its function is in the dimeric enzyme. Interestingly, this lobe was modelled into SDH - yet SDH contains no structural zinc ion. There is the possibility that the tetrameric long chain alcohol dehydrogenases polymerise in a similar manner to that found in GDH.

Region in monomer A	Sites of interaction
Structural lobe (Arg 102,Asn 108)	Monomer D loop (Arg 173,Asp 179)
Structural lobe (Asn 108,Ser 110,Ile 111)	Monomer D loop (Asn267,Asn 268)
Structural lobe (Ile 99,Arg 102,Ile 103)	Monomer D loop (Arg 296)
Structural lobe (Gly 104)	Monomer B structural lobe (Ile 99,Arg 102,Ile 103)
Structural lobe (Cys 109)	Monomer C loop (Arg 173)
Loop (Arg 173, Asp 179)	Monomer D structural lobe (Asn102, Arg108)
Loop (Phe 177,Asp 180,Ser 181)	Monomer C helix (Lys 308,Ile 309)
Loop (Asp 179)	Monomer C structural lobe (Gly104)
Loop (Asn 267,Asn 268)	Monomer D structural lobe (Asn108,Ser 110,Ile 111)
Loop (Arg 296)	Monomer D structural lobe (Cys109)
Loop (Asp144, Glu146)	Monomer C helix (Ala 328)
Helix (Ile 309,Arg 323,Arg 324)	Monomer C interconnecting helix (Asp 144,Glu146)
Nucleotide-binding domain	Monomer D nucleotide binding domain

Table 7.2 The monomers in GDH are designated A, B, C and D. The AD dimer and the BC dimer are equivalent to the LADH dimer. Residues that participated in the subunit/subunit interactions were determined using CONTACTS . Numbers in brackets indicate the numbering of the residues which interact.

Multiple sequence alignments

The structural homology of GDH to LADH (and the models) has enabled a structure/function relationship, and a phylogenetic relationship between GDH and members of the Zn-ADH family to be explored.

The purpose of multiple sequence alignments are to

- Provide a framework from which other sequences can be modelled.
- To aid in determining functionally important residues and try to ascertain the reasons for the functional differences between these enzymes.
- To allow a phylogenetic relationship between these enzymes.

Each of these features shall be discussed.

Structural framework

The structure of GDH and LADH were superposed using SHP (D. Stuart, unpublished program), which in conjunction with the SDH and YADH models were used to create a structurally based sequence alignment within the GCG package. This sequence alignment served as a block file, with structurally equivalent regions within this alignment designated as block regions. The block file was then used to align other members of the Zn-ADH family present in

1 10 REGION 1 20 30 REGION 2 40

Adhe_Horse	S	T	A	G	K	V	I	K	C	K	A	A	V	L	W	E	E	K	K	P	F	S	L	E	E	V	E	V	A	P	P	P	K	A	H	E	V	R								
Dhso_Sheep					A	K	P	A	A	E	N	L	S	L	V	V	H	G	P	G	D	L	R	L	E	N	Y	P	I	P	P	P	K	G	P	N	E	V	L							
Adh1_Yeast					S	I	P	E	T	E	Q	K	A	I	V	T	D	A	P	K	G	G	L	E	Y	K	T	T	I	D	M	P	E	K	P	E	H	Y	D	A	K					
Gdh_therm								T	E	Q	K	A	I	V	T	D	A	P	K	G	G	L	V	K	Y	T	T	I	D	M	P	E	K	P	E	H	Y	D	A	K						
adh_aspni					M	C	I	P	T	M	Q	W	A	Q	V	A	E	K	V	G	G	P	L	F	V	Y	K	Q	I	P	V	P	K	P	P	G	P	K	A	H	E	V	L			
ADH_chicken	S	T	V	G	K	V	I	K	C	K	A	A	V	L	W	E	A	N	K	P	L	F	S	L	E	E	V	E	V	A	P	P	P	K	A	H	E	V	L							
ADH_Cl.					M	M	R	F	T	L	P	R	D	I	Y	Y	G	K	G	S	E	Q	L	K	N	L	K	G	K	K	A	M	L	V	L	G	G									
ADH_aspergillus					M	C	I	P	T	M	Q	W	A	Q	V	A	E	K	V	G	G	P	L	L	T	V	Y	K	Q	I	P	V	P	K	P	P	G	P	K	A	H	E	V	L		
ADH_barley	M	A	T	A	G	K	V	I	K	C	K	A	A	V	A	W	E	A	G	K	P	L	L	T	M	E	K	E	D	I	P	V	P	K	P	P	K	A	H	E	V	L				
ADH_kl_yeast					M	A	A	S	I	P	E	T	E	Q	K	G	V	I	F	Y	E	N	G	G	E	P	L	L	S	Y	K	E	D	I	P	V	P	K	P	P	K	A	H	E	V	L
ADH_maize					M	A	T	A	G	K	V	I	K	C	K	A	A	V	A	W	E	A	G	K	P	L	L	S	I	E	E	E	V	E	V	A	P	P	P	K	A	H	E	V	L	
ADH_rice					M	A	T	A	G	K	V	I	K	C	K	A	A	V	A	W	E	A	G	K	P	L	L	S	I	E	E	E	V	E	V	A	P	P	P	K	A	H	E	V	L	
ADH_pea	M	S	N	T	V	G	Q	I	I	K	C	R	A	A	V	A	W	E	A	G	K	P	L	L	S	I	E	E	E	V	E	V	A	P	P	P	K	A	H	E	V	L				
ADH_pearl					M	A	T	A	G	K	V	I	K	C	K	A	A	V	A	W	E	A	G	K	P	L	L	S	I	E	E	E	V	E	V	A	P	P	P	K	A	H	E	V	L	
ADH_petunia	M	S	N	T	A	G	Q	V	I	R	C	K	A	A	V	A	W	E	A	G	K	P	L	L	S	I	E	E	E	V	E	V	A	P	P	P	K	A	H	E	V	L				
ADH_frog					S	T	A	G	K	V	I	K	C	K	A	A	V	A	W	E	A	G	K	P	L	L	S	I	E	E	E	V	E	V	A	P	P	P	K	A	H	E	V	L		
ADH_potato	M	S	T	T	V	G	Q	V	I	R	C	K	A	A	V	A	W	E	A	G	K	P	L	L	V	I	M	E	E	E	V	E	V	A	P	P	P	K	A	H	E	V	L			
ADH_clover	M	S	N	T	A	G	Q	V	I	K	C	R	A	A	V	A	W	E	A	G	K	P	L	L	V	I	M	E	E	E	V	E	V	A	P	P	P	K	A	H	E	V	L			
ADH_Zymomonas										M	K	A	A	V	I	T	K	D	H	T	L	I	E	V	M	E	E	V	E	V	A	P	P	P	K	A	H	E	V	L						
ADH_tomato	M	S	T	T	V	G	Q	V	I	R	C	K	A	A	V	A	W	E	A	G	K	P	L	F	S	L	E	E	V	E	V	A	P	P	P	K	A	H	E	V	L					
ADH_quail					S	T	A	G	K	V	I	K	C	K	A	A	V	L	W	E	A	N	K	P	L	F	S	L	E	E	V	E	V	A	P	P	P	K	A	H	E	V	L			
ADH_human	M	S	T	T	G	Q	V	I	R	C	K	A	A	I	L	W	K	L	G	A	P	P	F	F	S	L	E	E	E	V	E	V	A	P	P	P	K	A	H	E	V	L				
ADH_mouse					S	T	A	G	K	V	I	K	C	K	A	A	V	L	W	E	L	H	K	P	F	F	S	L	E	E	E	V	E	V	A	P	P	P	K	A	H	E	V	L		
ADH_rat					A	N	Q	V	I	R	C	K	A	A	V	A	W	E	A	G	K	P	L	L	S	I	E	E	E	V	E	V	A	P	P	P	K	A	H	E	V	L				
ADH_alcaligenes					M	T	A	M	M	K	A	A	V	F	V	E	P	G	R	I	L	E	L	A	D	K	P	I	P	P	D	I	G	S	Y	N	D	A	I							
ADH_clobe										M	K	G	F	A	M	L	G	I	N	K	L	L	G	W	I	E	K	E	G	I	P	V	A	G	S	Y	N	D	A	I						
ADH_Strawberry	M	S	S	T	E	G	K	V	I	C	C	R	A	A	V	A	W	E	A	G	K	P	L	L	V	I	E	E	E	V	E	V	A	P	P	P	K	A	H	E	V	L				
ADH_cod					A	T	V	G	K	V	I	K	C	K	A	A	V	A	W	E	A	N	K	P	L	L	V	I	E	E	E	V	E	V	A	P	P	P	K	A	H	E	V	L		
ADH_macaque					S	T	A	G	K	V	I	K	C	K	A	A	V	L	W	E	V	M	K	P	L	L	F	S	L	E	E	V	E	V	A	P	P	P	K	A	H	E	V	L		
ADH_Mycobacterium									M	S	T	V	A	A	A	A	M	S	A	T	E	P	L	L	F	S	L	E	E	V	E	V	A	P	P	P	K	A	H	E	V	L				
ADH_baboon	S	T	A	G	K	V	I	K	C	K	A	A	V	A	W	E	A	G	K	P	L	L	F	S	L	E	E	V	E	V	A	P	P	P	K	A	H	E	V	L						
ADH_Schiz.					M	T	I	P	D	K	Q	L	G	A	V	F	H	T	H	G	G	P	E	N	V	K	F	E	E	E	V	V	P	V	A	E	P	G	P	D	E	V	L			
ADH_thermophile					M	K	G	F	A	M	L	S	I	A	K	V	G	W	I	L	E	K	E	K	E	K	P	A	P	G	P	D	E	V	L											
ADH_Sulfolobus										M	R	A	V	R	L	V	E	I	G	K	P	L	S	L	Q	E	I	G	V	P	K	P	K	G	P	D	E	V	L							
SDH_silk									M	T	E	N	Y	S	A	V	L	H	G	A	N	D	L	V	R	L	E	K	I	P	V	P	E	I	N	D	D	E	V	L						
SDH_human					A	A	A	K	P	N	N	L	S	L	V	V	H	G	P	G	D	L	R	L	E	N	Y	P	I	P	E	L	G	P	N	D	V	L								
SDH_rat	M	A	A	P	A	K	G	E	N	L	S	L	V	V	H	G	P	G	D	I	R	L	E	N	Y	P	I	P	E	L	G	P	N	D	V	L										

REGION 3 50 60 70 80 REGION 4

Adhe_Horse	I	K	M	V	A	T	G	I	C	R	S	D	D	H	V	V	S	G	T	L	V	T	P	L	P	V	I	A	G	H	E	A	A			
Dhso_Sheep	L	K	M	H	S	V	G	I	C	G	S	D	D	V	H	Y	W	Q	G	R	I	D	F	V	V	K	K	P	M	V	L	G	H	E	A	A
Adh1_Yeast	I	N	V	K	Y	S	G	V	C	H	T	D	L	H	A	W	H	G	D	P	L	P	T	K	L	P	L	V	G	G	H	E	A	A		
Gdh_therm	L	S	P	V	Y	I	G	I	C	C	S	T	D	R	G	E	V	A	G	A	N	P	E	G	E	N	F	L	V	L	G	H	E	A	A	
adh_aspni	V	K	I	R	Y	S	G	V	C	H	T	D	L	H	A	M	M	G	H	P	I	P	V	K	M	P	L	V	G	G	H	E	A	A		
ADH_chicken	I	K	I	V	A	T	G	I	C	R	S	D	D	H	V	V	T	G	A	L	A	M	P	F	P	I	I	L	G	H	E	A	A			
ADH_Cl.	G	S	M	K	R	F	G	F	Y	D	K	V	L	G	V	L	K	E	A	G	I	E	V	K	L	I	E	G	V	E	P	P	P	S	P	
ADH_aspergillus	V	K	I	R	Y	S	G	V	C	H	T	D	L	H	A	M	M	G	H	P	I	P	V	K	M	P	L	V	G	G	H	E	A	A		
ADH_barley	V	K	I	L	F	T	S	L	C	H	T	D	V	Y	F	W	E	A	K	Q	Q	I	P	V	K	M	P	P	R	I	F	G	H	E	A	A
ADH_kl_yeast	I	N	V	K	Y	S	G	V	C	H	T	D	L	H	A	W	K	G	D	P	L	P	T	K	L	P	L	V	G	G	H	E	A	A		
ADH_maize	V	K	I	L	F	T	S	L	C	H	T	D	V	Y	F	W	E	A	K	G	Q	T	P	V	F	P	P	R	I	F	G	H	E	A	A	
ADH_rice	V	K	I	L	F	T	S	L	C	H	T	D	V	Y	F	W	E	A	K	G	Q	T	P	V	F	P	P	R	I	F	G	H	E	A	A	
ADH_pea	L	K	I	L	F	T	S	L	C	H	T	D	V	Y	F	W	E	A	K	G	Q	T	P	L	F	P	P	R	I	F	G	H	E	A	A	
ADH_pearl	V	K	I	L	Y	T	S	L	C	H	T	D	V	Y	F	W	E	A	K	G	Q	T	P	L	F	P	P	R	I	F	G	H	E	A	A	
ADH_petunia	L	K	I	L	F	T	S	L	C	H	T	D	V	Y	F	W	E	A	K	G	Q	T	P	L	F	P	P	R	I	F	G	H	E	A	A	
ADH_frog	V	K	I	V	A	T	G	I	C	R	S	D	D	H	V	I	S	G	A	L	S	D	M	K	F	P	P	V	I	L	G	H	E	A	A	
ADH_potato	L	K	I	L	Y	T	S	L	C	H	T	D	V	Y	F	W	E	A	K	G	Q	N	P	V	F	P	P	R	I	L	G	H	E	A	A	
ADH_clover	L	K	I	L	F	T	S	L	C	H	T	D	V	Y	F	W	E	A	K	G	Q	T	P	L	F	P	P	R	I	F	G	H	E	A	A	
ADH_Zymomonas	L	E	M	E	Y	C	G	V	C	H	T	D	L	H	V	K	N	G	D	F	G	D	E	T	G	R	I	T	T	G	H	E	A	A		
ADH_tomato	L	K	I	L	Y	T	S	L	C	H	T	D	V	Y	F	W	E	A	K	G	Q	N	P	V	F	P	P	R	I	L	G	H	E</			

Adhe_Horse	QNLS	MNPM	LLLL	SG	RTWK	GAI	F	GGF	KSK	DS	VPK	LVA	DFM
Dhso_Sheep	TSS	VP	LVHAA	TRE	V	DI	K	GVF	RYC	NT	WPM	AI	SMLA
Adh1_Yeast	KCCS	D	VFN	QVVK	S	S	I	S	I	V	GS	Y	VGN
Gdh_therm	PGYP	VDG	EDI	DYI	VE	RNI	TI	AGS	V	DA	AK	I	H
adh_aspni		YLK	APV	I	NT	V	VR	M	T	K	GS	Y	VGN
ADH_chicken	QKIS	FDP	M	L	I	F	S	G	R	T	W	K	G
ADH_Cl.	S	L	A	G	N	T	N	E	L	V	D	S	L
ADH_aspergillus		YLK	APV	I	NT	V	VR	M	T	K	GS	Y	VGN
ADH_barley	AEFK	T	H	P	M	N	F	L	N	E	R	T	L
ADH_kl_yeast	KC	K	S	D	V	F	N	Q	V	V	K	S	I
ADH_maize	AEFK	T	H	P	M	N	F	L	N	E	R	T	L
ADH_rice	AEFK	T	H	P	M	N	F	L	N	E	R	T	L
ADH_pea	DAFK	T	H	P	M	N	F	L	N	E	R	T	L
ADH_pearl	AEFK	T	H	P	M	N	F	L	N	E	R	T	L
ADH_petunia	DAFK	T	H	P	M	N	F	L	N	E	R	T	L
ADH_frog	AVMS	F	D	P	L	L	I	L	T	G	R	I	L
ADH_potato	AVFK	T	H	P	M	N	F	L	N	E	R	T	L
ADH_clover	DAFK	T	H	P	M	N	F	L	N	E	R	T	L
ADH_Zymomonas	MDL	S	I	P	R	L	V	L	N	D	G	I	E
ADH_tomato	AVFK	T	H	P	M	N	F	L	N	E	R	T	L
ADH_quail	QKIS	F	D	P	M	L	I	F	S	G	R	T	W
ADH_human	VQLK	I	S	G	Q	L	F	F	S	G	R	S	L
ADH_mouse	QNLS	MNPM	LLLL	LL	G	R	T	W	K	G	A	I	F
ADH_rat	EEIS	T	R	P	F	Q	L	V	T	G	R	T	W
ADH_alcaligenes	TIPR	S	A	F	A	A	G	L	V	T	G	R	T
ADH_clobe	ALLIPR	V	E	W	G	C	G	M	A	H	K	T	I
ADH_Strawberry	AVFT	T	H	P	M	N	F	L	N	E	R	T	L
ADH_cod	DVA	T	R	P	I	Q	L	I	A	G	R	T	W
ADH_macaque	QNLS	I	N	P	M	L	L	T	G	R	T	W	K
ADH_Mycobacterium	MAVP	A	F	A	L	A	L	M	R	S	L	A	G
ADH_baboon	QNLS	I	N	P	M	L	L	T	G	R	T	W	K
ADH_Schiz.	KL	G	A	D	I	F	W	L	T	V	K	M	L
ADH_thermophile	VLVPR	L	E	W	G	C	G	M	A	H	K	T	I
ADH_Sulfolobus	LHY	H	A	P	L	I	T	S	E	I	Q	F	V
SDH_silk	VE	L	P	L	S	O	A	L	R	E	V	D	V
SDH_human	T	S	V	L	P	L	H	A	A	T	R	E	V
SDH_rat	I	N	L	P	L	V	H	A	A	V	R	E	V

Adhe_Horse	AK	K	F	A	L	D	P	L	I	T	H	V	L	P
Dhso_Sheep	SK	S	V	N	V	K	K	S	P	I	T	H	R	V
Adh1_Yeast	AR	G	L	I	K	S	P	I	K	V	V	G	S	
Gdh_therm	NWNR	R	H	P	D	A	M	K	S	I	I	T	Y	
adh_aspni	AK	G	L	I	K	S	P	I	K	V	V	G	S	
ADH_chicken	KK	K	F	V	L	D	P	L	I	T	H	V	L	
ADH_Cl.	I	H	E	Q	E	F	K	N	K	V	D	L	I	
ADH_aspergillus	AR	G	L	I	K	S	P	I	K	V	V	G	S	
ADH_barley	KK	E	L	E	V	E	K	F	I	T	H	S	V	
ADH_kl_yeast	SR	E	L	E	V	E	K	F	I	T	H	S	V	
ADH_maize	KK	E	L	E	V	E	K	F	I	T	H	S	V	
ADH_rice	KK	E	L	E	V	E	K	F	I	T	H	S	V	
ADH_pea	KG	E	L	E	V	E	K	F	I	T	H	S	V	
ADH_pearl	KK	E	L	E	V	E	K	F	I	T	H	S	V	
ADH_petunia	KK	E	L	E	V	E	K	F	I	T	H	S	V	
ADH_frog	NK	K	F	L	D	P	L	I	T	H	V	L	P	
ADH_potato	NK	K	F	L	D	P	L	I	T	H	V	L	P	
ADH_clover	KG	E	L	E	V	E	K	F	I	T	H	S	V	
ADH_Zymomonas	AE	E	L	E	V	E	K	F	I	T	H	S	V	
ADH_tomato	NK	E	L	E	V	E	K	F	I	T	H	S	V	
ADH_quail	KK	K	F	V	L	D	P	L	I	T	H	V	L	
ADH_human	AE	K	L	N	L	D	P	L	I	T	H	V	L	
ADH_mouse	AK	K	F	P	L	D	P	L	I	T	H	V	L	
ADH_rat	SK	K	I	K	V	D	L	P	L	I	T	H	V	
ADH_alcaligenes	SG	R	V	D	L	G	A	L	V	T	H	Q	Y	
ADH_clobe	R	V	D	L	S	K	L	V	T	H	Q	Y	P	
ADH_Strawberry	NK	E	L	E	V	D	K	F	I	T	H	V	L	
ADH_cod	DK	K	V	K	L	D	E	F	I	T	H	R	M	
ADH_macaque	AK	K	F	S	L	D	A	L	I	T	H	V	L	
ADH_Mycobacterium	A	E	H	G	V	T	P	E	I	E	L	I	E	
ADH_baboon	AK	K	F	L	D	A	L	I	T	H	V	L	P	
ADH_Schiz.	SR	S	L	V	K	P	Y	Y	K	V	Q	P	R	
ADH_thermophile	YK	R	V	D	P	S	K	M	L	V	T	H	R	
ADH_Sulfolobus	VK	P	M	I	T	K	T	M	K	G	F	D	N	
SDH_silk	SG	A	I	P	L	D	K	F	I	T	H	R	F	
SDH_human	SK	S	V	N	V	K	P	L	V	T	H	R	F	
SDH_rat	SK	T	L	N	V	K	P	L	V	T	H	R	F	

Adhe_Horse	L T F
Dhso_Sheep	I K C D P S D Q N P
Adh1_Yeast	V D T S K
Gdh_therm	
adh_aspni	L E M P E
ADH_chicken	L V L
ADH_Cl.	C V Y Y G T E V D F
ADH_aspergillus	L E M P E
ADH_barley	I R M D N
ADH_kl_yeast	V D T S K
ADH_maize	I R M E N
ADH_rice	I R M E N
ADH_pea	I K M E E
ADH_pearl	I R M E N
ADH_petunia	I T M E H
ADH_frog	L T F
ADH_potato	I T M E D
ADH_clover	I R M E E
ADH_Zymomonas	V D F T H H
ADH_tomato	I T M A D
ADH_quail	L V L
ADH_human	
ADH_mouse	L T F
ADH_rat	L K L
ADH_alcaligenes	A I K P H
ADH_clobe	V I L
ADH_Strawberry	I T M E E
ADH_cod	L S L E
ADH_macaque	L T F
ADH_Mycobacterium	I D I S A L
ADH_baboon	L T F
ADH_Schiz.	L D L S K
ADH_thermophile	V I L A
ADH_Sulfolobus	L I P
SDH_silk	I H V Q N
SDH_human	L K C D P S D Q N P
SDH_rat	I K C D P N D Q N P

Fig. 7.14 Sequence alignment of the Zn-ADH family. The first four sequences in bold represent the block sequences. The boxed regions represent the block regions. The region in italics represent the nucleotide binding domain. ADH from *Clostridium acetobutylicum* (seq 7) is reported to have ADH type activity, but from the sequence alignment it is clear that it shares no resemblance to the Zn-ADH family. Shaded regions represent the strictly conserved residues. Abbreviations used in the alignment (from top to bottom) (All enzymes are ADH, unless stated): *Equus caballus*, *Ovis aries* sorbitol dehydrogenase (SDH), *Saccharomyces cerevisiae*, glucose dehydrogenase from *Thermoplasma acidophilum*, *Aspergillus nidulans*, *Gallus gallus*, *Clostridium acetobutylicum*, *Hordeum vulgare*, *Kluyveromyces lactis*, *Zea mays*, *Oryza sativa*, *Pisum sativum*, *Pennisetum americanum*, *Petunia hybrida*, *Rana perezi*, *Solanum tuberosum*, *Trifolium repens*, *Zymomonas mobilis*, tomato, *Coturnix japonica*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Alcaligenes eutrophus*, *Clostridium beijerinckii*, *Fragaria ananassa*, *Gadus callarius*, *Macaca mulatta*, *Mycobacterium bovis*, *Papio hamadrysa*, *Schizosaccharomyces pombe*, *Thermoanaerobium brockii*, *Sulfolobus solfataricus*, *Bombyx mori* SDH, *Homo sapiens* SDH and *Rattus norvegicus* SDH.

the Swissprot data base (37 different species in total) using the AMPS package (Barton & Sternberg, 1990). A heavy weighting function was applied to minimise insertions/deletions within the block regions (table 7.3). The sequence alignment was displayed using ALSCRIPT (fig. 7.14) (Barton, 1993).

In order to provide a structural framework from which other sequences can be modelled, at least one structure of an enzyme within the alignment needs to be solved. Plausible models will aid in determining the structural framework, however these models are biased towards the structure by the very nature of their creation. The

Block region	GDH secondary structure	ADH secondary structure	No. of amino acids
1	$\beta 1, \beta 2$	$\beta III:1$	6
2	$\beta 3, \beta 4$	$\beta III:2$	6
3	$\beta 5, \alpha 1$	$\beta II:1, \beta I:1, \alpha 1$	21
4	$\beta 6, \beta 7$	$\beta III:3, \beta I:2, \beta II:2$	17
5	$\beta 8, \beta 9$	$\beta I:3$	10
6	$\beta 11, \beta 12$	$\beta II:4, \beta I:4$	24
7	$\alpha 2$	$\alpha 2$	10
8	βA	βA	8
9	αB	αB	16
10	βB	βB	10
11	αC	αC	6
12	βC	βC	4
13	βD	βD	22
14	βE	βE	11
15	αF	αF	4
16	$\alpha 3$	$\alpha 3$	12

Table 7.3 The secondary structural elements of GDH and LADH that comprise the block regions.

structure solution of GDH has revealed extensive structural homology to LADH. Two independent observations of essentially the same structure, from evolutionary divergent species, has enabled a structural blueprint of the Zn-ADH family to be

formulated, as defined by the block regions. The regions that were homologous within the block sequence were weighted to maintain the conservation of the structurally equivalent residues. In effect, this forces most of the insertion/deletions to occur within loops, or areas that are structurally divergent falling between the conserved regions, which are occupied mainly by secondary structural elements. The structural regions containing and surrounding the ligands to the catalytic zinc appear to be very well conserved.

The sequence alignment displays which residues of an enzyme are structurally equivalent to those from another enzyme. When the residues occur within a structurally homologous region this assignment is more confident, conversely the likelihood of structural equivalence is low throughout the entire Zn-ADH family outside these regions.

Multiple sequence alignments of the Zn-ADH family have previously been reported. Jornvall (1987) compared 17 different members (11 species) using pairwise alignment, which was based on the secondary structural elements of LADH. Later, with the arrival of more sequences, Kim & Plapp (1992) compared 47 members (31 species) using a progressive sequence alignment, but discarded the structural bias of the LADH structure in favour of an objective alignment.

Conservation of structure, and hence function is a much greater driving force than conservation of the primary sequence. The problem associated with the progressive sequence alignment is that it will try to maximise the local identity of a region of the sequence - regardless of where the insertion/deletion of the residues would occur. The structural consequences of the putative insertion/deletion are ignored. For example,

many of the dimeric Zn-ADHs possess a 21 amino acid surface loop. The tetrameric enzymes are believed not to contain this surface loop. There is no evidence to suggest that the tetrameric and dimeric enzymes share structural homology in this region, and thus a single gap in this alignment (Jornvall) is a true representation of the data as opposed to the 2 gap alignment of Kim & Plapp. The structure of GDH has shown that there is no region structurally homologous to the LADH surface loop. Another example of maintenance of local sequence identity in preference to structural integrity was in the assignment of residue 93 of yeast ADHs. Despite biochemical evidence that residue 93 should be a tryptophan (the Trp plays a role in determining the substrate specificity of the YADHS), Kim & Plapp suggest that this position may be occupied by Leu which would result an alteration of the β strand preceding this loop.

Structure-function relationships

Jornvall found that there were 22 strictly conserved residues throughout the family, whereas the more comprehensive study by Kim & Plapp revealed that there were only 9 strictly conserved residues. In this study, the number of strictly conserved residues are seven (and this falls to four if ξ -crystallin is included in the alignment) The position of these in the alignment are Gly 80, 100, 201, 271, Cys 51, His 81 and Glu 82. Four of these strictly conserved residues are glycine, which is indicative of dissimilar sequences, whereas the remainder are involved in ligation to the catalytic zinc. Comparison of evolutionary distant sequences, such as the Archaeal and Eucarya/Bacterial enzymes may give the minimal number of strictly conserved residues within the Zn-ADH family. In addition to the strictly conserved residues, there are

conserved¹ and consensus residues that are present throughout the majority of the Zn-ADH family, as indicated by Kim & Plapp. GDH shares and differs from some of these conserved/consensus residues, with some of these non-conservative changes occurring within structurally equivalent regions (see table 7.4). GDH residues putatively relating to substrate binding are discussed in Chapter 8.

Position 52: In LADH, this Arg (Arg 47) interacts with NAD. The His residue present in a number of Zn-ADHs is also believed to interact with NAD. However, this interaction doesn't appear to be essential, as a few Zn-ADH members possess a Gly, and GDH possesses a serine. Noticably, the 2 bacterial NADP-dependent Zn-ADHs possess a Thr at this position (a conservative change with respect to GDH), although the consequence of this observation is unknown.

Position 56: This His in LADH (His 51) plays an important role in the proton relay system of the LADH mechanism (Ehrig *et al.*, 1991). Some Zn-ADHs have a Tyr, that probably cannot act as a proton acceptor, whereas GDH possesses a Gly which definitely cannot interact in the proton relay system, which suggests that the GDH mechanism must differ in some respects to that of LADH. Furthermore, there is no apparent residue lining the active site that can replace the function of His 51.

Some of the changes (76, 91, 94), which are far removed from the active site, can be attributed to GDH possessing or interacting with loops that have a different conformation to that of LADH, and thus there is no requirement to conserve these residues.

¹ within a given alignment, a residue is deemed conserved if the majority (>80%) of the residues are identical at a specified position.

Position 112: In LADH, Cys(97) acts as a ligand to the structural zinc. Although GDH possesses a structural zinc, the corresponding lobe is structurally dissimilar to that of LADH. The fourth ligand to the structural zinc was found to be Asp 115 in GDH position 131 in the alignment).

N°.	Conserved/ consensus residue	Corresponding GDH residue	N°.	Conserved/ consensus residue	Corresponding GDH residue
36	P	P	223	G	G*
40	E	D	227	A	V*
51	C	C*	228	V/I	V*
52	G/H/R	S*	230	G	G*
53	S/T	T*	231	L/A	S*
54	D	D*	233	G	G*
56	H/Y	G*	234	G/A/P	S*
76	P	F*	235	V/L/I	E*
80	G	G*	236	G	A*
81	H	H*	239	V/A	Y*
82	E	E*	244	K/R	V*
85	G	G*	247	A/G	F*
87	V	V*	250	R	D*
91	G	R*	251	I/V	V*
94	V	G	252	I/V/L	T*
100	G	G*	253	G/A	M*
101	D	D*	254	V/I	V*
103	V/A	V*	255	D	N*
112	C	P	263	K/R	K*
115	C	C	271	G	G
118	C	C	325	G	G*
126	C	C	330	V	F*
161	S/G	G	331	G	G*
163	F	M	332	V/L	T*
178	K/H/R	K*	406	F/Y	S
179	I/V/L	V*	413	G	G
188	V/A	A*	418	R/K	R
193	C	E*			
197	T	N*			

Table 7.4 The conserved and consensus residues of the Zn-ADH family in relation to GDH. *

denotes a residue within a block region. Numbering as defined in the alignment.

Positions 230-236: relates to the GxGxxG/A sequence motif. The Ala has been defined as a diagnostic for NADP binding proteins, such as GRS and GDH. However, other NADP binding Zn-ADHs do not possess this Ala, which is consistent with the theme of Baker *et al.* (1992) that the terminal Gly/Ala residue of the motif does not determine coenzyme specificity, but determines the hydrogen-bonding system between the adenine ribose and the GxGxxG/A loop.

Positions 255-259: a negatively charged and positively charged environment at the end of β B (position 255-259) plays an important role in conferring NAD and NADP specificity respectively. Asp 223 (position 255) of LADH confers NAD specificity, whereas the equivalent residue in GDH is Asn, which would enable a weak interaction with the 2'OH of the adenine ribose, and thus explain why GDH exhibits activity towards NAD (see chapter 8). Zn-ADH from *Clostridium*, *Mycobacterium* and *Thermoanaerobium* possess a Gly or Ser at this position, which could not interact with NAD. His 217 (position 257) of GDH interacts with the 2' phosphate of NADP. *Thermoanaerobium* and *Clostridium* Zn-ADH possess Arg at this position, whereas *Mycobacterium* possess a Lys at position 259.

Positions 197, 239, 330: of GDH contains residues that interact with NAD(P), which in LADH do not interact with the coenzyme.

Position 244: of LADH is occupied by a solvent exposed positively charged residue. IN GDH (and the *Mycobacterium*), this position is occupied by a hydrophobic residue situated within a hydrophobic core. Monomer C of GDH (see table 7.2 for definition

of quaternary structure) shelters this region from the solvent, and hence provides a hydrophobic environment for the Val.

Position 250: is a site of compensated exchange between GDH and LADH. In LADH (Arg 218) and GDH (Asp 210), this residue occupying this position interacts with the an oppositely charged residue at position 274 (Arg-Glu 239 in LADH, Asp-Lys 233 in GDH). In addition, Asp 210 in GDH is stabilised by an additional interaction in GDH (by Arg 187, position 225). Interestingly, a similar feature is observed in *Mycobacterium* and the SDHs.

Phylogenetic concerns

The purpose of the phylogenetic analysis was not to reanalyse the grouping of the Zn-ADH family, but merely to observe where GDH resides phylogenetically within the family. The tree was constructed using the PHYLIP package (Appendix 1, Phylogenetic tree) (fig. 7.15).

The animal, plant and yeast Zn-ADHs group together respectively, in accordance with previous phylogenetic studies (Kim & Plapp, 1992). GDH was observed to lie with the tetrameric polyol dehydrogenases, and lie on the same fork of the tree as the the NADP requiring Zn-ADHs. Interestingly, GDH did not group with the Archaeal Zn-ADH from *Sulfolobus solfataricus* which was situated next to the *Mycobacterium* at the base of the yeast Zn-ADH fork. Many more Archaeal and Bacterial Zn-ADH sequences are required before conclusions regarding Woese's Archaeal tree can be drawn.

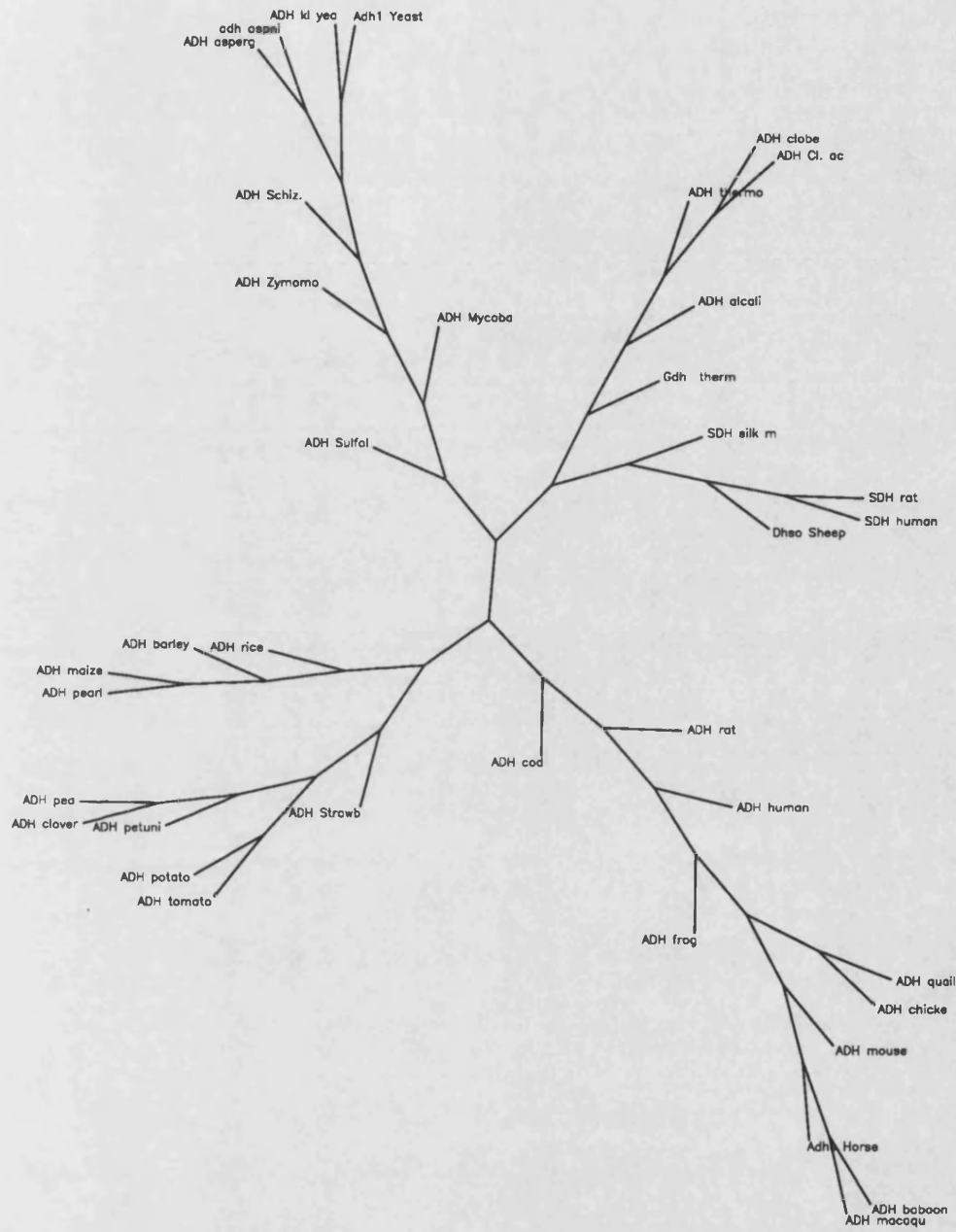


Fig. 7.15 An unrooted phylogenetic tree of the Zn-ADH family. Abbreviations as defined in fig.

7.14

Chapter 8

COENZYME AND SUBSTRATE INTERACTIONS

Introduction

Glucose dehydrogenase was given the status of a dual-cofactor specific enzyme. However on the basis of the published K_m values for NADP and NAD (0.11mM, >30mM respectively) , it should be stated that GDH preferentially utilises NADP (Bright *et al.*, 1993; Smith *et al.*, 1989). An investigation concerning the basis for the observed coenzyme specificity was undertaken. The structural determinants involved in coenzyme specificity were discussed in chapter 1. In addition, the nature of the substrate binding site was explored.

Nucleotide interactions

From the structurally based sequence alignment of the category 1 $\beta\alpha\beta$ fold (Fig. 8.1), it can be seen that GDH belongs to this category, and that it possesses the GxGxxA sequence motif which is indicative of some NADP binding proteins.

The mode of nucleotide binding (in particular that of the adenosine moiety) is structurally conserved throughout the dehydrogenase family (Wierenga *et al.*, 1985). Structures of

observed dual-cofactor specificity. Asn 215 of GDH (position 42 in Fig. 8.1; Fig. 8.2) is structurally equivalent to the Asp 223 of LADH and Arg 218 of glutathione reductase. The presence of Asn 215 would enable a weak interaction to be made with the adenosine 2' hydroxyl - this would explain the weak binding of NAD to GDH. The binding of NADP requires stabilisation of the adenosine 2' phosphate. The hydrophilic character at the end of β B and the following loop would aid in this stabilisation. In addition, specific interactions between the 2' phosphate and GDH appear likely. His 217 (position 45 in Fig. 8.1) is in good hydrogen bonding distance to the 2' phosphate of NADP.

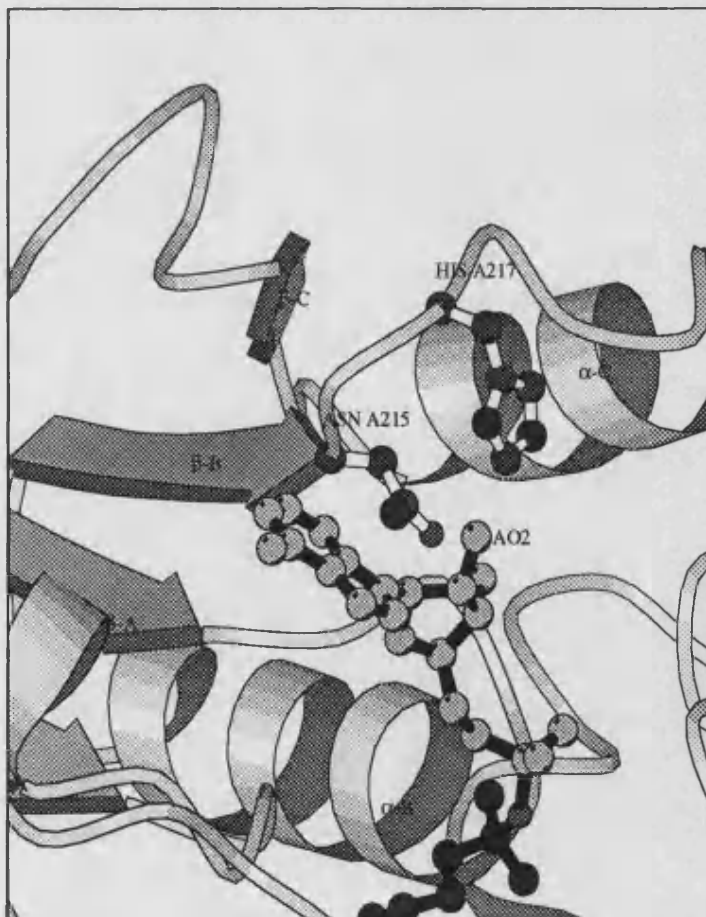


Fig. 8.2 Schematic representation of the proposed binding of the adenine-ribose moiety of NAD.

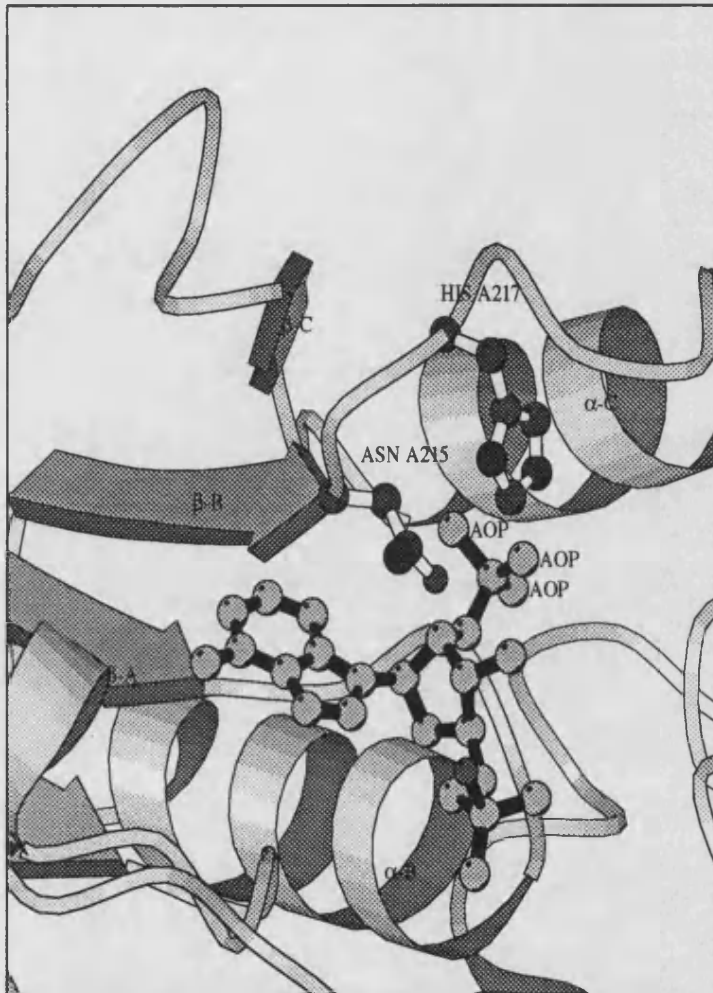


Fig. 8.3 Schematic representation of the proposed binding of the adeneine ribose moiety of NADP.

The optimum pH for GDH activity is pH 7; Glu 219, which interacts with His 217, or alternatively the incoming phosphate moiety of NADP, could raise the pKa of His 217, enabling the His 217 to interact with the adenosine 2' phosphate of NADP. These putative strong interactions between NADP and GDH may explain the preferential binding to NADP (fig. 8.3).

Coenzyme and substrate interactions

NAD	LADH	GDH	GDH/NAD interaction	
Adenine moiety	Phe 198	Ile 191	hydrophobic	
	Val 222	Val 214	hydrophobic	
	Asp 223	Asn 215	H-bond to AN9	
	Ile 224	Arg 216	hydrophobic	
	Pro 243	Tyr 237	aromatic interaction	
	Ile 250			
	Ile 269			
	Thr 274	Thr 257	H-bond to AN2	
	Arg 271	Asp 254	H-bond to AN6 & AN1	
		Phe 261	aromatic interaction	
		His 217	<i>aromatic</i> interaction	
	Adenine ribose	Gly 199	Gly 192	hydrophobic
		Asp 223	Asn 215	H-bond to AO-2 & AO-3
Lys 228		Thr 220	H-bond to AO-3	
Ile 200		Ser 193	H-bond to AO-3	
		His 217	H-bond to 2'P of NADP	
		N-193	H-bond to AO-3	
Pyrophosphate	N-202	N-195	H-bond	
	N203	N-196	H-bond	
	Arg 47			
		Glu 196	H-bond?	
	Arg 369	Arg 352	charge interaction	
	C=O 269	C=O 252	H-bond to NO-3	
Nicotinamide ribose	N-294	N-276	H-bond to NO-2	
	His 51			
	Gly 270	Gly 253	hydrophobic	
	Val 203	Glu 196	H-bond to NO-5	
		Thr 251	H-bond to NO-4	
	Gly 293	Gly 275	hydrophobic	
	Val 294	Thr 276	H-bond to NO-2	
	Nicotinamide	Val 203		
		Val 292	Phe 274	aromatic interaction
		Val 294	Thr 276	
Thr 178		Asn 159	H-bond to NO-7	
	N-319	C=O 303	H-bond to carboxamide	

Table 8.1 Comparison of the NAD(P) interactions with LADH and GDH. The interactions between cofactor and GDH were determined using CONTACTS. Interactions within 3.6Å were regarded as potential hydrogen bonding /hydrophobic interactions. The contacts between GDH/NAD and GDH/NADP were essentially the same, except that the 2'phosphate of NADP could interact with His 217. For description of NAD(P) nomenclature used, see Appendix 1, Substrate/NAD structures.

From the modelled GDH/NAD(P) complex, it was observed that a few of the GDH residues (notably Tyr 237) interacted too close with the NAD(P). It was uncertain whether the proximity was due to an error in the co-ordinates, or a slightly incorrect positioning of the NAD(P).

To reduce any bad interactions, the enzyme-cofactor model was minimised(see appendix 1, minimisation). The deviation of the minimised NAD(P) (fig. 8.4) with respect to the superposed NAD(P) was slight, resulting in a change of some of the putative interactions listed in table 8.1. A small movement of some of the side chains of GDH within the cofactor binding site was also observed. The most interesting change of NADP featured the adenine ribose moiety. Tyr 237, Phe 261 and Ile 191, must impose some restraints in the positioning of the adenine ring. The optimal positioning of the adenine ring (represented by the minimised model) caused the ribose moiety to change its orientation with respect to Asn 215, such that AO-3 (of NAD) could form a putative strong hydrogen bond with Asn 215, whilst AO-2 (of NAD) was too distant from Asn 215. However, the 2'phosphate of the minimised NADP was still within good bonding distance of His 217. Naturally, these interactions are only putative. A crystallographic study of the GDH/coenzyme (analogue) complexes would reveal the exact nature of the interactions. Crystallographic investigations of enzyme/coenzyme analogue complexes would entail crystal soaking experiments. To this end, a preliminary low resolution (6Å) Fo-Fc map of the GDH/ATP-ribose complex has been obtained (data not shown)(soak conditions: 1mM ATP-ribose /2 hours). The analogue bound crystal was isomorphous to the native cell. The position of the difference density was in the vicinity of the GxGxxA region, although

the modelled adenine moiety of this ATP-analogue did not directly superpose onto this density. Further studies are necessary.

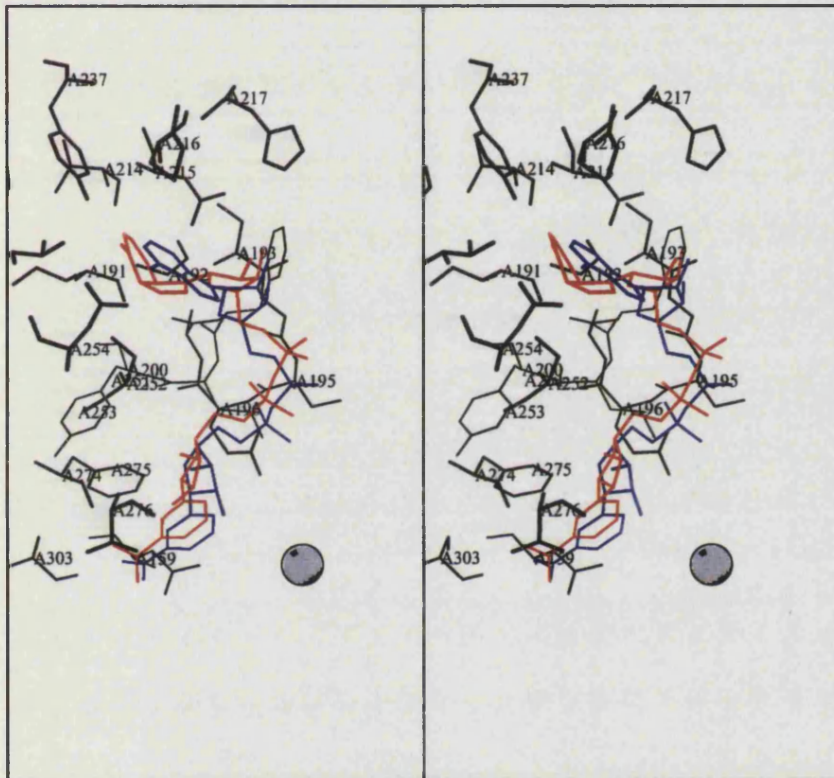


Fig. 8.4 Stereoview of the minimized (red) and superposed NAD (blue) and the associated ligands.

The GxGxxG/A loop

It has been observed that the pyrophosphate moiety of NAD(P) is not optimally positioned to form good hydrogen bonds with any of the main chain atoms within this loop (Wierenga *et al.*, 1985). In addition, the GxGxxG/A region is essentially hydrophobic (see

fig. 7.12, fig. 8.1.). Instead, the pyrophosphate moiety has been found to interact with the helix dipole of α B in a number of dehydrogenases.

In GDH, the GxGxxA region is GSGSEA. From all the corresponding sequence motifs to date, GDH is one of the few sequences possessing a charged residue (the glutamate) within this loop (G6PDH also possesses a charged residue, M. Adams, personal comms).

There are two putative consequences of this glutamate

- the glutamate would negate the helix dipole of α B, and thus remove the pyrophosphate/helix dipole mode of binding.
- The glutamate interacts with the pyrophosphate NAD(P) (see fig. 8.5)

Although this latter interaction sounds dubious, carboxyl-pyrophosphate interactions have been documented previously (Baker *et al.*, 1992). For the Glu to interact with the pyrophosphate moiety, either the Glu (very unlikely) or the pyrophosphate needs to be in the basic form. NADP specific and dual cofactor Glutamate dehydrogenases possess a negatively charged residue at the end of β B, which contradicts the general coenzyme-specificity determining rule (namely a positively charged residue at the end of β B confers NADP specificity). Baker *et al* have hypothesised that the negatively charged residue is able to interact with the 2' phosphate of NADP when the phosphate moiety is in the monobasic or dibasic form.

The presence of the Glu 196 would lower the pK value of the pyrophosphate, resulting in the pyrophosphate oxygen atom being protonated (and thus able to hydrogen bond).

It was also noticed that the carboxyl group of Glu 196 would interact favourably with the delocalised positive charge on the nicotinamide ring.

The hydrogen-bonding network

Baker *et al.* (1992) have identified a hydrogen bonding scheme between the adenine ribose and the GxGxxG/A region. When the motif is GxGxxA the hydrogen bonding to the adenine ribose is direct - the main chain NH group of the second residue in the loop is pointing in such a way (because of the steric interactions imposed by the alanine residue) that it can form a direct hydrogen bond to an oxygen atom of the adenine ribose. In the GxGxxG motif, the hydrogen bonding scheme is indirect. The same NH group (as above) points in towards the space that would have been occupied by the

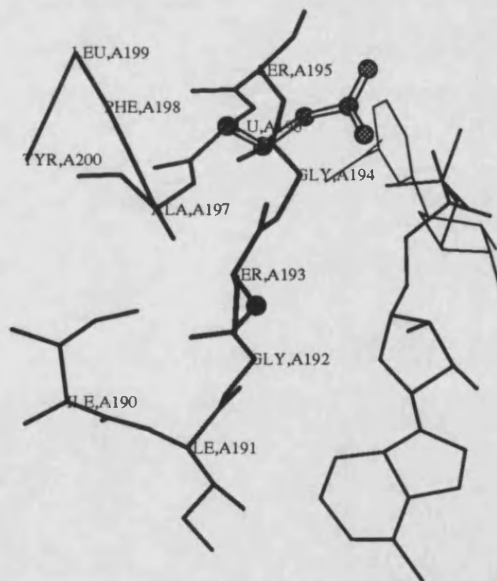


Fig. 8.5 The GDH GSGSEA loop interacting with NAD. The Glu 196 interacts with the nicotinamide ring and the pyrophosphate moiety. The NH-group of Ser 193 points towards the adenosine ribose.

alanine. The orientation of the NH group is such that it can hydrogen bond with the residue at the end of β B, which in turn hydrogen bonds to the adenine ribose.

GDH possesses the GxGxxA motif, and was found to exhibit the direct hydrogen bonding scheme (fig. 8.5), in accordance with the observations of Baker *et al.*

Substrate binding

Inhibitor/analogue complexes, coenzyme bound complexes and ternary complexes of LADH have been reported (Eklund & Brändén, 1979, and references therein). Hence the mode of substrate and coenzyme binding, and consequences are well known. From these studies, the differing substrate specificities exhibited by some members of the Zn-ADH family, in particular SDH, have been rationalised.

The catalytic mechanism of LADH proceeds via a compulsory order mechanism, in which NAD binds first, causes a conformational change, which then allows a productive mode of binding for the alcohol (Theorell & Chance, 1951). No new residues are brought into the active site by this conformational change- but serves to make the active site more hydrophobic so that hydride transfer can take place. In the apo- form, a water molecule co-ordinated to the catalytic zinc prevents the substrate from binding directly to the zinc (Eklund *et al.*, 1982). The conformational change results in the zinc-bound water ligand being excluded from the active site - thereby allowing the alcohol to bind.

GDH and LADH have homologous structures and catalyse similar reactions. Ergo, GDH is hypothesised to have a similar catalytic mechanism to that of LADH. This assumption allows the nature of the substrate binding site of GDH to be explored.

The question arises: where does GDH bind the glucose and what is the mode of glucose binding? The presence of a putative catalytic zinc in GDH suggests that glucose should be directly bound to this zinc. In addition, the glucose must be positioned in the vicinity of the nicotinamide ring of NAD(P) for hydride transfer to occur. The position of the modelled nicotinamide ring is also close to the catalytic zinc atom.

The major functional difference between GDH and LADH is that GDH selectively oxidises the C1 secondary alcohol of glucose, whereas the latter predominantly binds a range of primary alcohols (Sund & Theorell, 1963). Therefore, for GDH to accommodate glucose, the substrate binding site of GDH must differ in comparison to LADH. A comparative study between the LADH (and SDH) and GDH substrate binding sites should explain the differences in the enzymes specificity.

The initial docking

GDH catalyses the conversion of glucose to the gluconolactone, whereby a hydride C1 is abstracted from the C1 atom of glucose. Thus the C1 hydride must point towards the C4 atom of the nicotinamide ring.

A comparison of the residues lining the substrate binding site of LADH (and SDH), to the structurally equivalent residues (where applicable) in GDH should highlight the similarities and dissimilarities between these enzymes (table 8.2). LADH has a hydrophobic substrate binding pocket, and despite sorbitol being a polyol, the SDH substrate binding pocket is also essentially hydrophobic (Eklund *et al.*, 1985).

LADH	GDH	LADH	GDH
Zn	Zn	Leu 141	not equivalent
Cys 46	Cys 40	Phe 140	not equivalent
His 67	His	Ser 48	Thr 42
Cys 174	Glu 155	Val 294	Thr 276
Leu 116	not equivalent	Leu 57	Leu 51
Phe 93	Val 92	Ile 318	Val 304

Table 8.2 The LADH ligands lining the substrate binding site, and the equivalent GDH residues

Leu 116, Phe 140 and Leu 141 are found within the surface loop of LADH, and thus accounts for GDH not possessing residues that are structurally equivalent. Despite the differences, there are some similarities between the binding sites, namely the catalytic zinc and the associated ligands. The consensus ligand Ser/Thr 48, which is present throughout the Zn-ADH family is present in GDH (Thr 42) (position 53 in table 7.4). In LADH and SDH, Ser/Thr 48 ligates to the substrate hydroxyl moiety which is bound to the carbon atom from which the proton is abstracted. This observation also points towards the C1 hydroxyl moiety of glucose being hydrogen bonded to the Thr 42 of GDH.

The model of SDH, a polyol dehydrogenase, greatly aided the further positioning of glucose within GDH (a polyol dehydrogenase). One of the notable features of the SDH model is that Glu 174 is a ligand to the catalytic zinc, contrasting with the conserved Cys 174 ligand present throughout the majority of the Zn-ADH family (position 193 in table 7.4). Glu 155 in GDH, which is structurally equivalent to Cys 174 of LADH and Glu 174 of SDH, is a ligand to the catalytic zinc. This Cys/Glu ligand is located within the

interconnecting helix of LADH and GDH respectively. LADH contains an extra residue preceding this ligand with respect to the GDH structure and SDH model. In SDH, this gap-Glu alignment provides sufficient space for the primary hydroxyl group, with the Glu ligating to this primary hydroxyl group, enabling the zinc to ligate to the C-2 hydroxyl. GDH contains only one primary hydroxyl group, at the C-6 position, thus based on the SDH model, the plausible scenario is that the primary C-6 hydroxyl of glucose is ligated to the Glu155 of GDH. Eklund *et al* (1985) pointed out that the Phe 93→Pro 93 mutation in SDH provided enough space for the positioning of the primary hydroxyl group in sorbitol - the Phe 93 residue in LADH would result in unfavourable steric clashes. Val 92 of GDH is the structurally equivalent residue to Phe 93 of LADH (position 107 in fig. 7.14). This smaller side chain in GDH would also accommodate the C-6 primary hydroxyl group of glucose.

These putative interactions restrict the glucose within the active site. Sorbitol is a linear/extended sugar, whereas glucose forms a 6 membered pyranose ring. Thus the remaining sorbitol/SDH interactions are not likely to be present in the glucose/GDH complex. In SDH and LADH, the hydroxyl group of the C2 atom (from which the proton is abstracted) is ligated to the zinc. Due to the ring form of glucose, there is an intervening C5 and pyranose oxygen between the C1 hydroxyl and the C6 hydroxyl. In order to maintain a close contact between the C1 hydroxyl and the zinc atom, the glucose has to move deeper into the active site. As a consequence of this, there are steric clashes between the C6 hydroxyl and neighbouring side chains. The GDH/cofactor/substrate complex was minimised (see appendix 1, minimisation) to relieve any strain.

The minimised complex

Fig 8.6 shows that the minimised GDH complex did not deviate greatly from the GDH structure. The greatest difference between the unminimised and minimised structures was

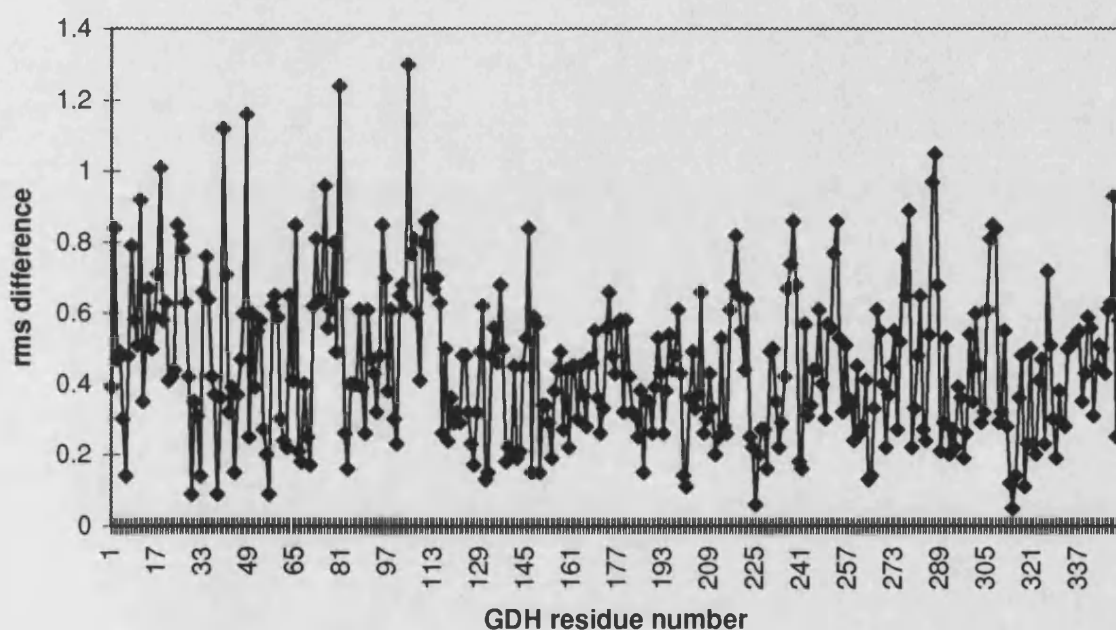


Fig. 8.6 rms difference between co-ordinated of the refined structure and the minimised complex. The greatest difference was seen at the sites of subunit interaction (100-115, 280-300) and crystal contacts (90-95).

found to correspond to the sites of subunit/subunit interaction and crystal contacts - which probably reflects that only a monomer was used in the minimisation. The nature of the active site did not change greatly.

Coenzyme and substrate interactions

From this minimised complex, the putative interactions concerning glucose were identified (fig. 8.7, table 8.3)

glucose	C1	C6	O1	O2	O3	O4	O5	O6
GDH/ NAD	NAD	NAD	Thr 42	Thr 276	Arg 93	Asp 305,	NAD	Glu155, Zn

Table 8.3 Interactions within 4Å were identified using contacts.

This table highlights that, because of the hydroxyls of glucose, there are no hydrophobic residues which interact closely with the glucose moiety. It is worthy to note that Thr 276 occurs within a block region (position 332, as defined in table 7.4). Throughout the rest of the Zn-ADH family, position 332 has Val/Leu as a consensus ligand. In LADH this residue points in towards the core of the protein, whereas in GDH the Thr points outwards towards the substrate binding pocket. This non-conservative change in GDH probably reflects the need for GDH to bind glucose. Arg 93 also occurs within a block region (position 108, region 5 in fig. 7.14); this position is not a consensus position within the Zn-ADH, however GDH is the only sequence possessing an arginine in this position. Asp 305 of GDH occurs within a region C-terminal to β F that is structurally dissimilar to that of LADH. In LADH, this region is occupied by two glycines (the hinge region) and points away from the active site. In GDH, this region continues to line the active site. Fig 8.8 represents a space filling model of glucose and NADP within the active site and displays how well the cofactor and glucose interact.

In this GDH/glucose complex model, the C1 hydroxyl did not directly bind to the zinc ion, although the C1-hydroxyl could be indirectly bound to the zinc

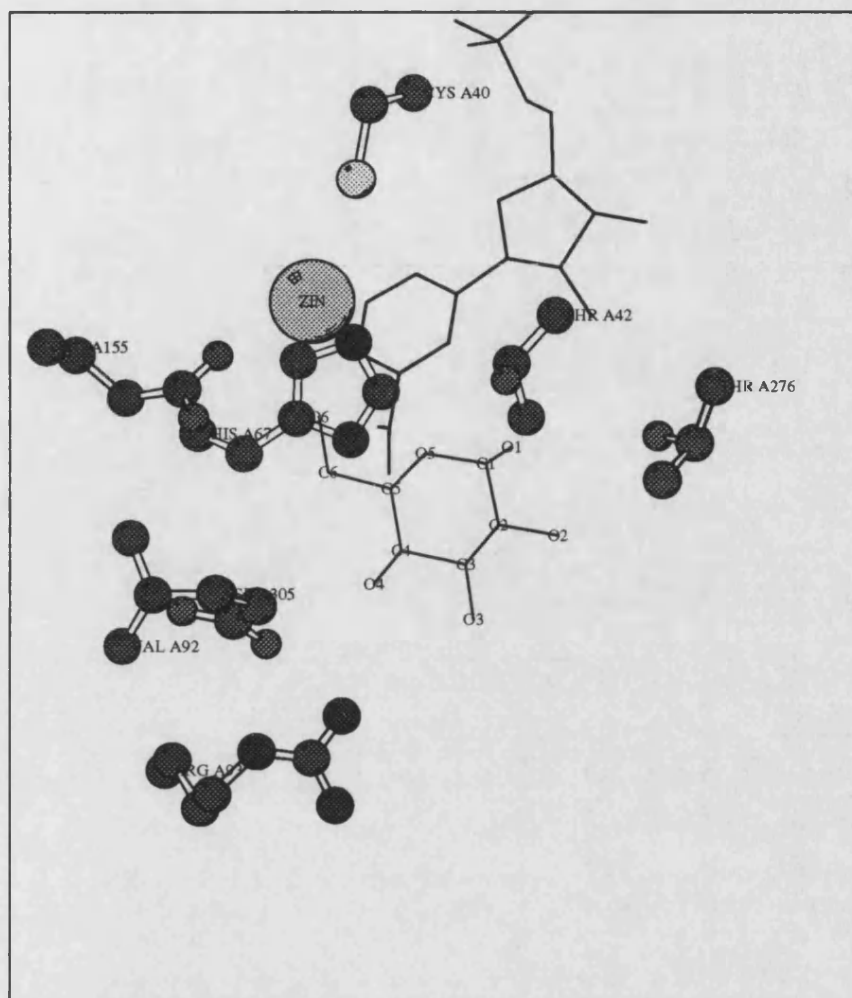


Fig. 8.7 Schematic representation of the putative ligands in the substrate binding site.

via a water molecule. A tentative hypothesis regarding this observation is discussed in Chapter 10. Other modes of glucose binding were explored, but no productive binding could be found - there were either too many steric clashes, or the C1 hydride of glucose was too distant from the NAD(P).

Substrate specificity

There have been previous kinetic investigations (Smith, 1989) into the substrate specificity of GDH (table 8.4) (for structures of sugars see appendix 1, substrate/nad structures).

The majority of this study was repeated and the results were in accordance with Smith.

Sugar	% rate relative to glucose	sugar	% rate relative to glucose
D-glucose	100	D-allose	7
D-Galactose	56	D-glucosamine	1
D-gulose	8	D-mannose	0.3
6-deoxy-D-glucose	10	D-altrose	0
2-deoxy-D-glucose	1	D-idose	0
D-xylose	8	D-ribose	0
D-fucose	16	Lactose	4
Sorbitol	0	ethanol	0

Table 8.4 Specificity of GDH towards a variety of sugars in relation to glucose.

To validate the substrate bound model of GDH, it should be possible to rationalise the observed substrate specificity from the structure. The 'glucose analogues' were superposed onto the modelled glucose. GDH displays a high activity towards galactose (a C4 epimer of glucose). Galactose can be easily accommodated into the substrate pocket, and although the C4 hydroxyl loses the close interaction with Asp 305, it is now able to interact with Arg 93 - which must serve to stabilise the galactose within the pocket. GDH exhibits little activity (<10%) with the majority of the sugars - which can be rationalised by steric clashes and loss of sugar/GDH stabilising bonds. However, GDH displays >10% activity towards fucose and 6-deoxy glucose, suggesting that the C6 hydroxyl group is important in binding to the Glu 155 and the zinc, but the loss of the hydroxyl group isn't

completely detrimental. A small hydrophobic group at C6 must be stabilised by the neighbouring hydrophobic Val 92.



Fig 8.8 Schematic representation of the active site displaying the space filling models of glucose and NADP.

GRASP

In addition to the putative specific interactions pertaining to the substrate binding site of GDH, the overall properties of the active site were explored, and compared to that of

LADH using GRASP (Nicholls, Bharadwaz, Honig - unpublished program).

Noticeably, there are more hydrophobic residues present in the LADH active site (fig. 8.12) than that found in GDH (fig. 8.10). However, a more prominent difference was observed in the surface potential of the respective active sites : GDH was markedly negative (fig. 8.9), whereas LADH was predominately positive (fig. 8.11). The functional implications of this observation are discussed in Chapter 10.

Continuing the modelling theme - the closed form

The mode of domain closure in LADH is well documented: upon binding of NAD, the catalytic domain of LADH undergoes a 10° rigid body rotation relative to the nucleotide binding domain. The hinge regions of this rotation are known (Colonna-Cesari *et al.*, 1986).

It is possible that GDH may undergo a similar conformational change upon cofactor binding. To this end, the catalytic domains of GDH were rotated by 10° relative to the nucleotide binding domain (using `O'). No obvious clashes within this monomeric closed form model were observed, however close intersubunit interactions were observed at the structural lobes of the monomers. This may indicate subtle differences in the domain closure of GDH with respect to LADH.

The ramifications of this model were not explored deeply - it was conceived to act as a possible molecular replacement search model when the closed form had been crystallised.



Fig. 8.9 The active site surface potential of GDH. Positive potential (blue), negative potential (red).

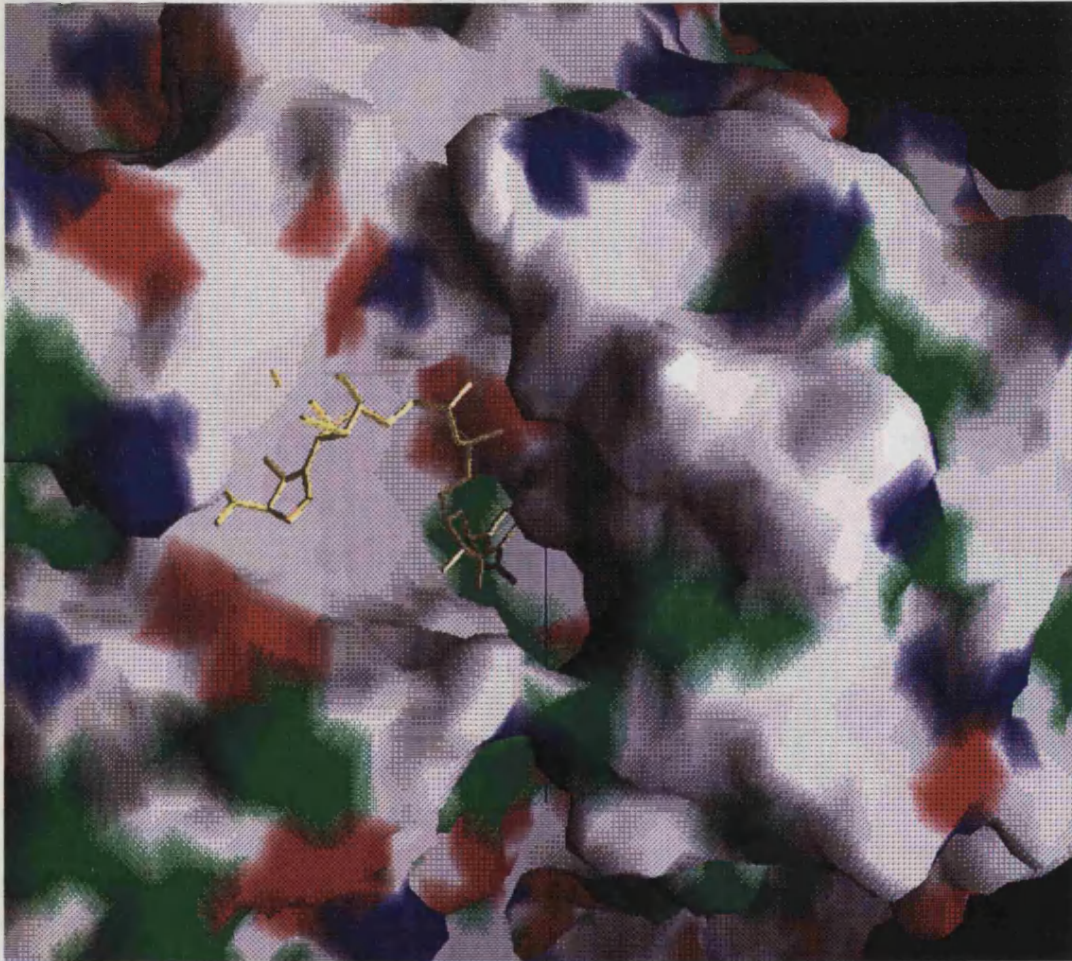


Fig. 8.10 The active site of GDH. Positive charge (blue), negative charge (red) and hydrophobics (green).



Fig. 8.11 The active site surface potential of LADH. Positive potential (blue), negative potential (red).

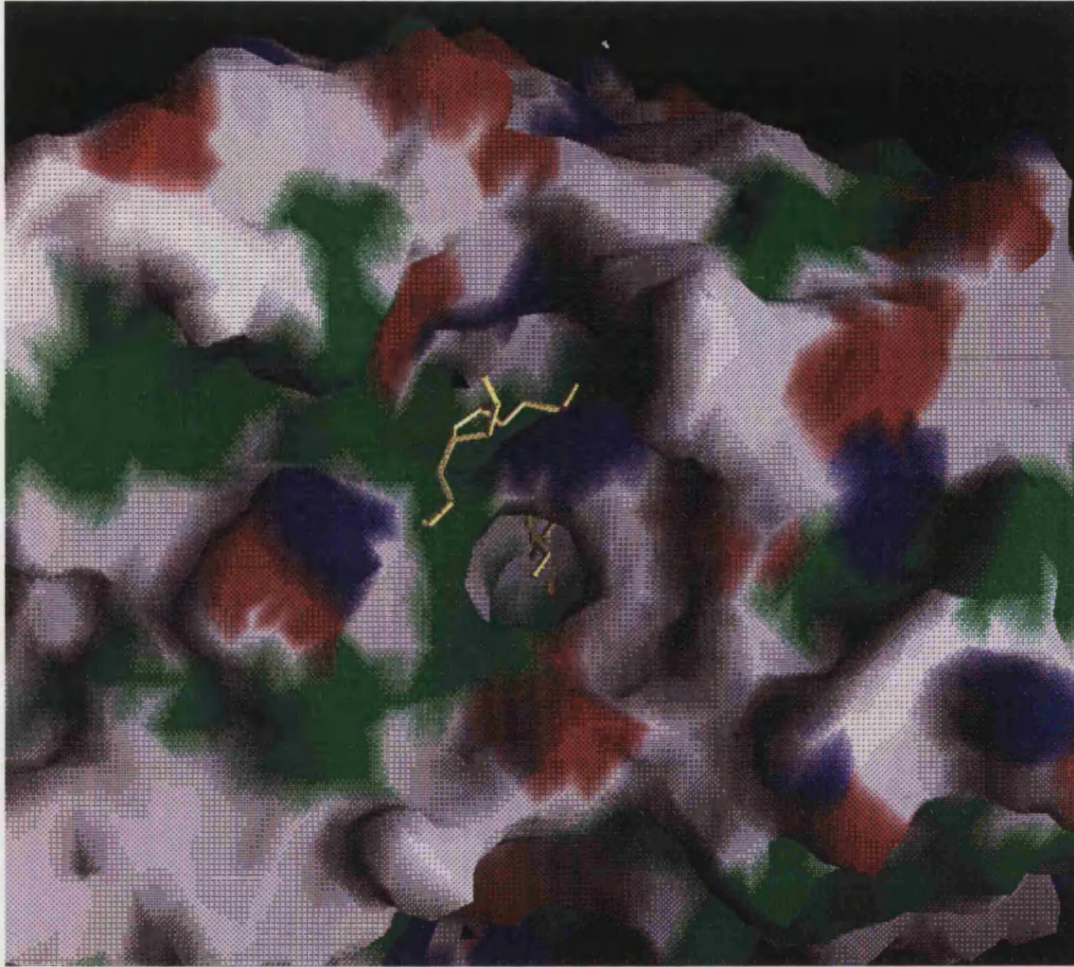


Fig. 8.12 The active site surface of LADH. Positive charge(blue), negative charge (red), and hydrophobic (green).

Chapter 9

THERMOSTABILITY

Introduction

The use of enzymes from thermophilic organisms in the biotechnology industry, or the increase in the intrinsic thermostability of mesophilic enzymes already used in industrial processes, has attracted much commercial interest (Hough & Danson, 1989). Hence there is interest in understanding, and being able to manipulate the factors that govern protein thermostability.

A detailed understanding of structural features conferring thermostability have arisen from

- Sequence and structural comparisons between thermophilic and mesophilic enzymes.
- Site directed mutation studies on proteins of known structure, such as the extensive studies of barnase and lysozyme (Fersht & Serrano, and references therein, 1993).
- Characterisation of the thermodynamics of the unfolded to folded transition (Jaenicke, 1991).

- Flexibility studies on proteins (Jaenicke, 1991).

Such thermophilic studies have been undertaken at a gross sequence level, for example the aliphatic index (Ikai, 1980), in an attempt to find an underlying difference in the amino acid compositional trend between thermophilic and mesophilic enzymes. These studies may act as a pointer towards understanding thermostable characteristics, however a prerequisite for the detailed understanding of the residual influences on protein thermostability is the knowledge of the three dimensional structure of the enzyme under investigation. However, the number of sequences determined outweigh the number of structures solved. With the assumption that there is an overall conservation of structure within a protein family, comparative studies within protein families - of which a representative structure is known - have highlighted putative thermophilic features (Menendez-Arias & Argos, 1989).

What features confer thermostability in enzymes? The hydrophobic effect is thought to be the principal driving force regarding protein stability. Strong and extensive hydrophobic interactions within the core of the protein implies a compact enzyme core in which there are few cavities. Cavity creating/filling mutations have been shown to decrease and increase the stability of the protein respectively (Kellis *et al.*, 1988; Lee & Levitt, 1991). Aromatic/aromatic interactions, which generally occur between secondary structural elements, are known to stabilise the conformation of the protein (Burley & Petsko, 1985). Helix stabilising factors have received much attention (Fersht & Serrano, 1993), although recent studies have also focused on β -sheet stabilising factors (Minor & Kim, 1994). Helix forming propensities of amino acids have been tabulated, where Ala and Gly are thought to be the most stabilising and

destabilising helical residues respectively. Reduction of helix flexibility and stabilisation of the helix dipole, via N and C terminal helix capping residues have also been shown to confer helical stability. Short β -branched side chains and aromatics are believed to be β -sheet stabilising residues.

Additional forces, for example introducing sites for ionic interactions and introducing disulphide bonds, have been shown to increase the thermostability of an enzyme - markedly so in the case of thymidylate synthase, where engineering of 2 disulphide bonds between two subunits led to an 20°C increase in thermostability (Gokhale *et al.*, 1994). Relatively few thermostability studies concerning oligomeric proteins have been reported, however strong subunit/subunit interactions are thought to stabilise the oligomeric state.

Engineering thermal stability into an enzyme is not straightforward though. The consequences of introducing a mutation into a protein is difficult to predict. In addition, mutations far removed from the active site may have drastic consequences on the activity of an enzyme. Nevertheless the stability and activity of site directed mutants within the protein core have been accurately predicted (Lee & Levitt, 1991).

Although the stability of an enzyme is believed to represent the difference in free energy between the native and unfolded state, a study has revealed that the stability of an enzyme can be predicted whilst completely neglecting the unfolded state (Lee & Levitt, 1991) .

GDH originates from a thermophilic Archaeon, but characterisation of the GDH thermostability, had not been examined. The principle reason for this was that GDH had no apparent homologue in the protein world and thus no thermophilic/mesophilic

comparisons could be made. This study has disclosed that GDH is approximately 80% structurally homologous to LADH. In addition, the nucleotide-binding domain is a conserved feature throughout a number of dehydrogenases. Mesophilic counterparts to the GDH structure have thus surfaced - enabling a comparative study to be made. As a result of the low sequence identity between GDH and the Zn-ADH family, it would be impossible to pin-point residue changes that confer thermostability. Instead, putative broad thermophilic features of GDH, have been highlighted. Although potential thermostable features have been observed, it must be remembered that GDH is from a moderate thermophile (55°C).

As well as GDH being a thermophilic enzyme, it is also an archaeal enzyme. Putative archaeal characteristics have also been proposed.

Compositional differences

There was no significant difference in either the frequency of the helix/strand stabilising amino acids or the amino acid composition between GDH and the members of the Zn-ADH family - with the exception of the cysteine content (table 9.1). It has been observed that thermophilic enzymes generally possess fewer cysteines than their mesophilic counterparts (Amaki *et al.*, 1994).

Throughout the Zn-ADH family, those members that contain 2 zinc ions/monomer generally possess at least 6 essential cysteines: 2 cysteines to the catalytic zinc, and 4 cysteines to the structural zinc. Numerous Zn-ADHs contain cysteines that do not participate in zinc co-ordination, whereas GDH only possesses cysteines that co-ordinate to the zinc. In addition, GDH has fewer cysteines that co-ordinate to the

zinc: one and three cysteines to the catalytic and structural zinc respectively (see chapter 7, The Zinc sites). Interestingly, the Zn-ADH from the thermophilic archaeon, *Sulfolobus solfataricus*, also only contains 3 cysteine ligands to the structural zinc. Thus the cysteine content of GDH certainly serves to strengthen the prior observation of Amaki *et al.* (1994)

species	No. of cysteines	species	No. of cysteines	species	No. of cysteines
Aspni	9	Petunia	13	Rat	14
Chicken	13	Frog	16	Alceu	9
Emeni	9	Potato	12	Strawb	13
Barley	13	Clover	13	Cod	14
Kl yeast	7	Zymmo	8	Macaque	15
Maize	13	Tomato	13	Mycob	8
Rice	13	Quail	13	Babbon	15
Pea	13	Human	16	Schizpombe	11
Pearlmillet	13	Mouse	15	Yeast	8
Sulfolobus	5	GDH	4	Horse	14

Table 9.1 Cysteine content of members of the Zn-ADH family (mean cysteine content $\approx 11.5 \sigma \approx$

3.0). For list of abbreviations see figure legend 7.14.

Stabilising helix capping residues

The type of residue at the amino and carboxy termini of helices are known to play a role in stabilising the helix (Richardson & Richardson, 1988). The order of preference of amino caps (determined for barnase) was shown to be Asp, Thr, Ser > Asn, Gly > Glu, Gln > His > Ala > Val (Fersht & Serrano, 1994). The amino cap hydrogen bond acceptors stabilise the helix by forming hydrogen bonds with the peptide backbone NH group of the N+3 residue. In addition, acidic residues interact favourably with the helix dipole. Glycine is by far the most favoured C-cap residue, followed by positively charged residues (which interact with the helix dipole). Of the structurally

homologous helices, it can be seen that GDH possesses stabilising N-cap residues (table 9.2). Many of these putative N-caps did not actually form hydrogen bonds to N+3 NH group, however a small rotation of these N-cap side chains could be visualised to form the stabilising hydrogen bond (which may be possible at the operational temperature of the enzyme). More definitive was the positioning of dipole-stabilising negatively charged residues in the vicinity of the N-cap residue within GDH. Fewer stabilising N-cap residues were present in the structurally homologous LADH helices (table 9.2). Thus from previous studies it appears that the N-cap helical residues of GDH appears to be a thermostable feature.

GDH			LADH		
helix	position	sequence	helix	position	sequence
α 1	41-47	S -T-D...	α 1	47-54	R -S-D...
α 2	154-174	T -E-P...	α A	175-184	G -F-S...
α B	195-205	G -S-E...	α B	202-213	G -G-V...
α C	221-228	T -E-N...	α C	229-235	K -F-A...
α E	255-262	D -P-T...	α E	272-280	R -L-D...
α 3	308-321	K -I-H...	α 3	324-336	K -D-S...
α 4	326-331	P -D-A...			

Table 9.2 The N-terminal helical residues of GDH and LADH. Residues in bold denote the start of the helix.

Aromatic/aromatic interactions

Fig. 9.1 represents a gallery of stereodiagrams highlighting the aromatic/aromatic interactions within a number of nucleotide-binding domains; table 9.3 displays the interacting aromatics for each nucleotide-binding domain. From Fig. 9.1 and table 9.3

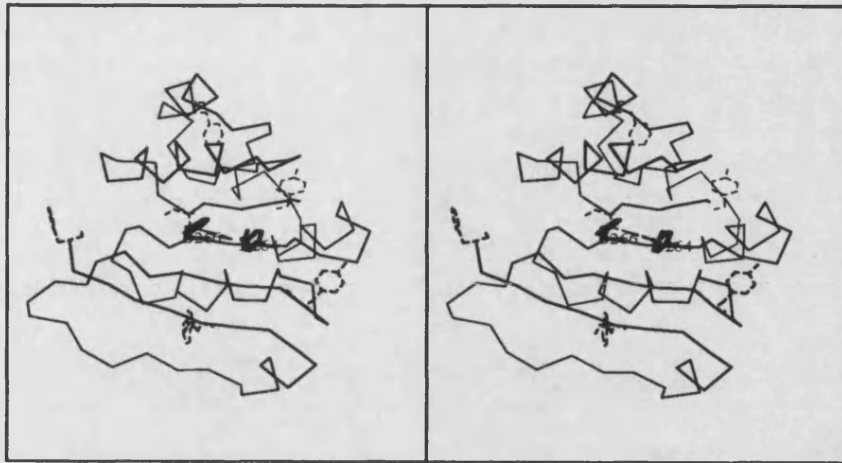
it can be seen that the number of aromatics within a nucleotide-binding domain is not a reliable marker for the number of aromatic/aromatic interactions.

As can be seen from the fig. 9.1 and table 9.3, GDH contains a multitude of aromatic interactions - which will undoubtedly serve to stabilise the nucleotide-binding domain.

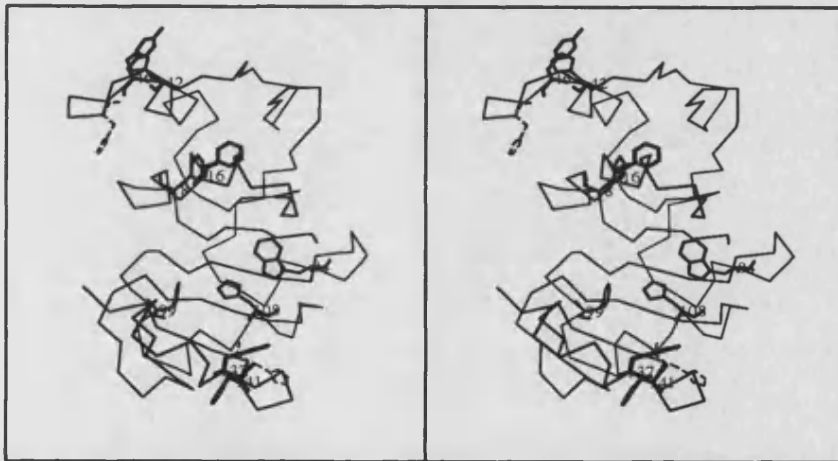
In addition, GDH possesses the smallest nucleotide-binding domain. Thus, the number of aromatic interactions in the GDH nucleotide-binding domain appears to be more significant when compared to the other nucleotide-binding domains. The nucleotide-binding domain of 2GPD also has an extensive network of aromatic

enzyme	No. of aromatics/ histidines	residues involved in aromatic interactions	length of domain (a.a.)
Glyceraldehyde 3- phosphate dehydrogenase (2GD1)	6/5	F8, F16, H42, Y46, W84, F99, H108, Y137, H141,	148
SDH	2/2	NONE	126
GDH	12/1	F198, Y200, F202, H217, F230, F234, Y237, F259, F261, F274, Y283, Y292.	114
LADH	8/0	F264, F266	126
glutathione reductase (3GRS)	6/2	F180, F181, Y197, H219, F226, F248	146
lactate dehydrogenase (4LDH)	6/4	Y85, , W150	142
Glyceraldehyde 3- phosphate dehydrogenase (2GPD)	14/2	F8, Y39, Y42, F44, Y46, F53, F74, W84, Y91, Y99, H108, F128, Y137	148
Phosphogluconate dehydrogenase (1PGD)	10/3	H24, F26, F31, H55, F83, Y136, W149, F156, W172	175

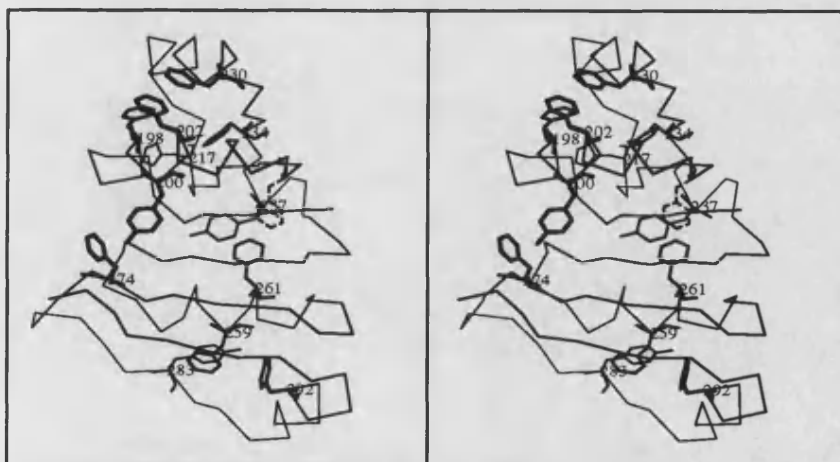
Table 9.3 An aromatic between 4.5Å and 7Å of another aromatic was ventured to participate in an aromatic interaction. 2GD1 is a thermophile.



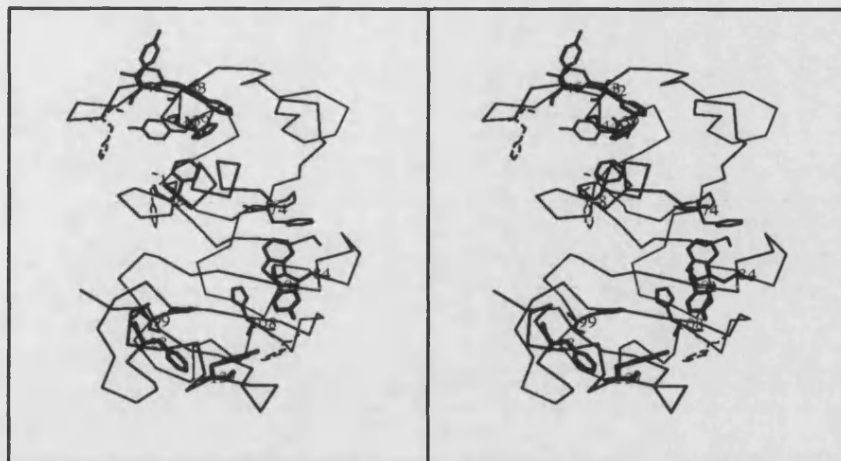
a)LADH



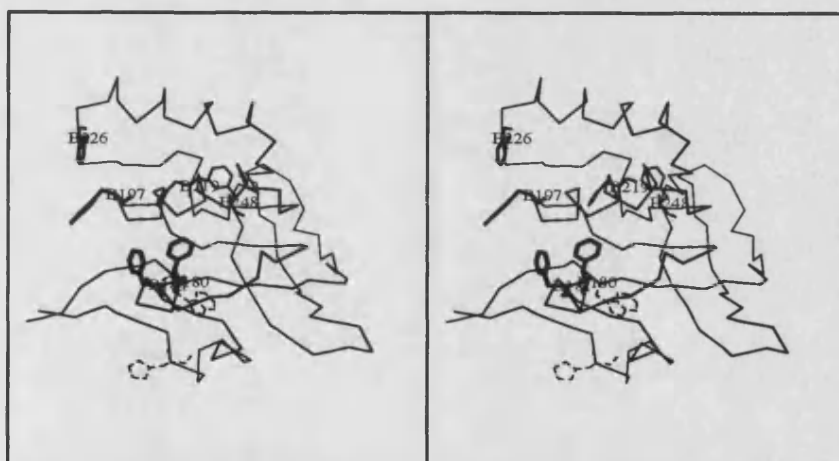
b)2GD1



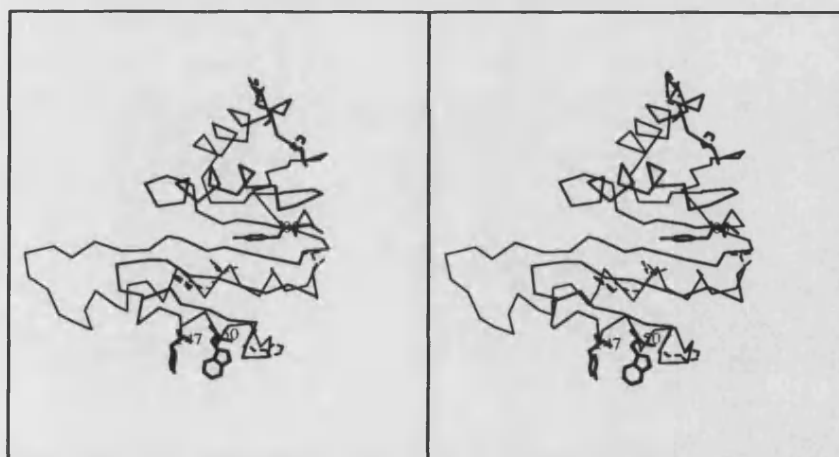
c)GDH



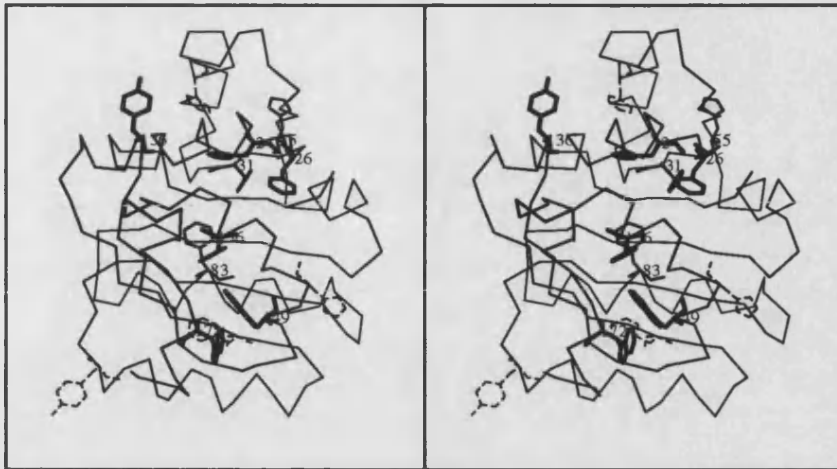
d)2GPD



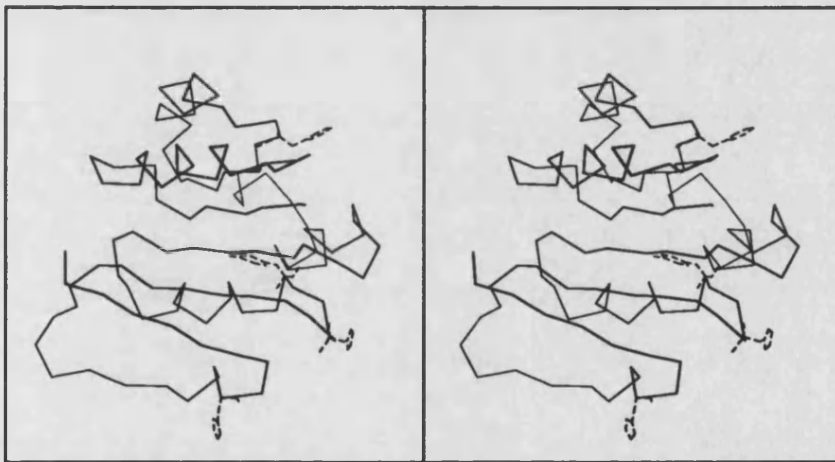
e)3GRS



f)4LDH



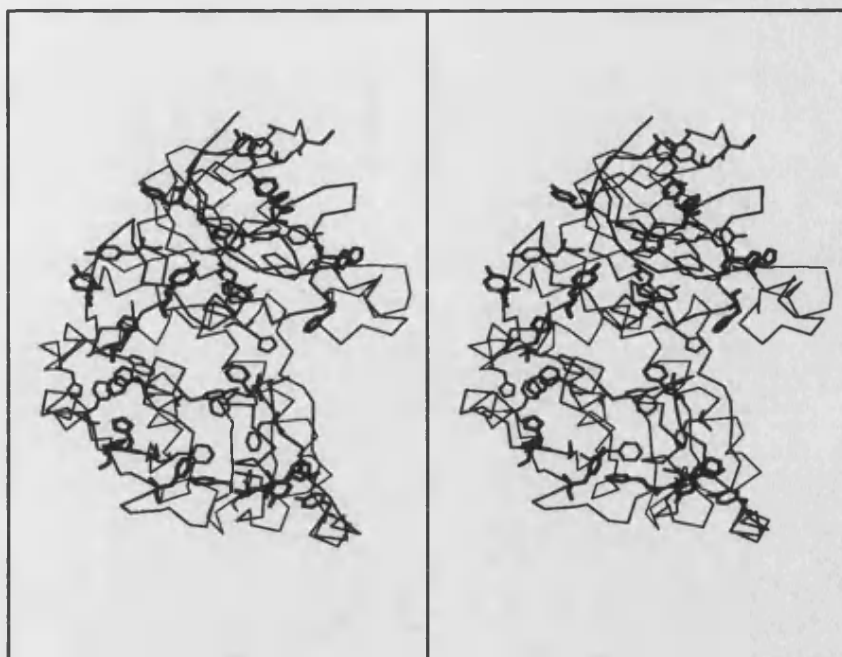
g)1PGD



h)SDH

Fig. 9.1 Backbone traces of the nucleotide-binding domains of a)LADH b) 2GD1 (thermophile) c) GDH d) 2GPD (mesophile) e) 3GRS f) 4LDH g) 1PGD h) SDH, displaying the aromatic/aromatic interactions. Aromatic and histidine residues that do not participate in aromatic/aromatic interactions are represented as dashed lines.

interactions (more so than the thermophilic counterpart), hence a large number of aromatic interactions isn't a characteristic restricted to thermophilic enzymes. The aromatic interactions within the entire GDH monomer and LADH monomer were also



a)



b)

Fig. 9.2 Aromatic residues within GDH (9.2a) and LADH (9.2b). GDH can be seen to contain more aromatic/aromatic interactions.

compared (see Fig 9.2).- GDH was observed to contain significantly more aromatic residues participating in aromatic interactions when compared to its mesophilic counterpart, which again must serve to stabilise the GDH structure.

The dimeric subunit/subunit interface

The nature of the intersubunit interface in GDH and LADH were investigated further, by closely examining the interactions involved (table 9.4, fig. 9.3, 9.4). The nature of the subunit interaction in LADH was reported to be predominantly hydrophobic, but in addition a large number (9) of intersubunit hydrogen bonds are present (approximately half of the total interactions). Despite the fact that GDH and LADH are structurally homologous in this region, the corresponding sequence identity is very low. As a result, there are very few homologous interactions between these enzymes.

In contrast to LADH, there are only 4 hydrogen bonding interactions between the GDH subunits. One reason for the lack of hydrogen bonding interactions is the coil region 279-285 in GDH (which is β S, 296-303 in LADH). The lack of defined secondary structure within 279-285 in GDH probably arises from two prolines and a glycine - which would disrupt the main chain hydrogen bonding. A more pronounced hydrophobic interaction was found in GDH - 80% of the interacting residues were involved in hydrophobic contacts (compared to 50% in LADH). Another interesting observation was that there were four aromatics interacting within this region in GDH (fig. 9.5) (compared to one tryptophan in LADH).

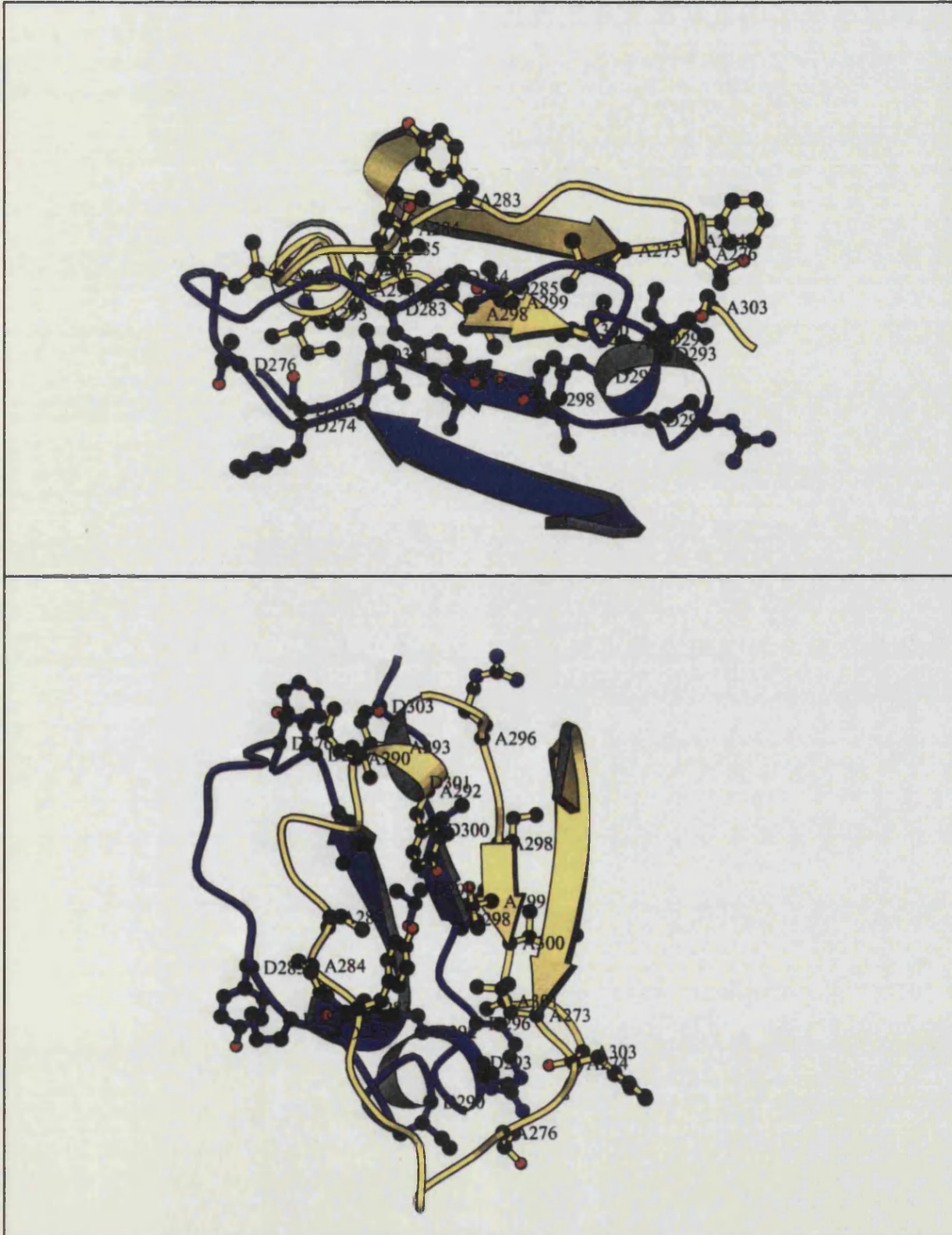


Fig. 9.4 Schematic representation of the nucleotide binding domain subunit interaction in GDH.

The residues participating in hydrophobic interactions are displayed as ball and stick objects.

Monomer A is in yellow, monomer D in blue.

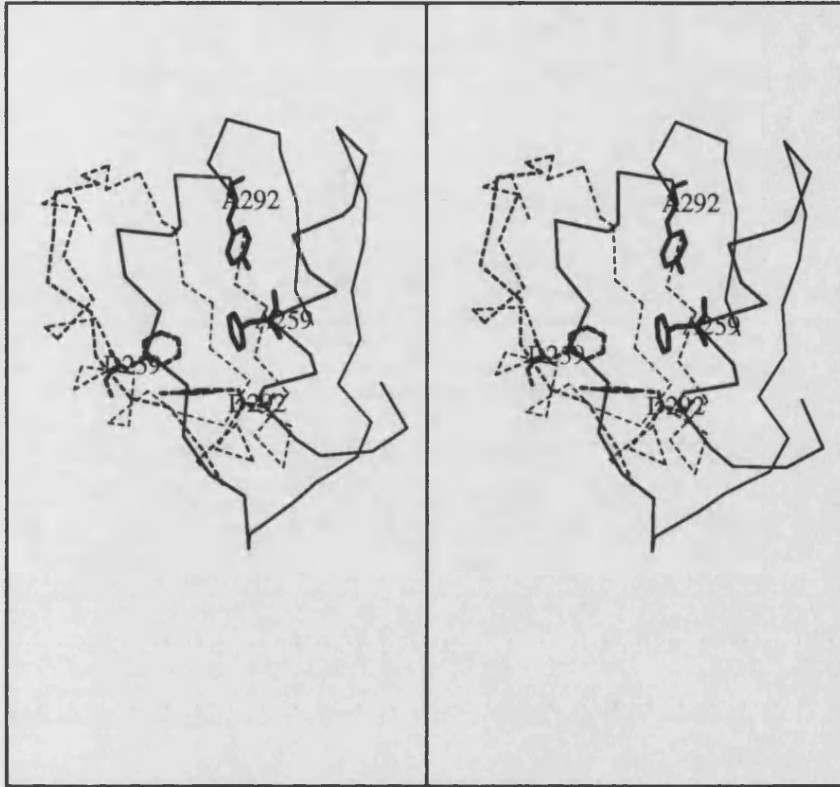


Fig. 9.5 The intersubunit aromatic interaction in GDH. Monomer D is represented as dashed lines.

Cavities

The presence of cavities within proteins is known to be a destabilising factor (Hubbard *et al.*, 1994), but nevertheless represent a small volume in relation to the protein ($\approx 2\%$). Generally, the larger the protein, the greater the number of cavities. The presence of cavities within proteins has proven to be independent of resolution and how well the structure is refined (Hubbard *et al.*, 1994). Thus a comparative study between GDH and LADH is valid. Cavities were detected and analysed using VOIDOO, following the protocol described by Kleywegt & Jones (1994), except that

a probe radius of 1.2\AA was used. The molecular volume of the GDH tetramer was determined to be $1.451 \times 10^5 \text{\AA}^3$, whilst that of the LADH dimer was $7.428 \times 10^4 \text{\AA}^3$. If GDH is considered as a dimer of LADH dimers, then it follows that GDH and LADH have very similar molecular volumes. Approximately 12-16 cavities were found in the GDH tetramer (total mean cavity volume of 202\AA^3), representing 3 cavities within the each monomer and one at each of the main subunit/subunit interfaces. The LADH dimer contained 4-7 cavities (total mean cavity volume of 89.5\AA^3), representing 3 cavities per monomer - with no cavities at the dimer interface. The % cavity volume (with respect to the protein) is 0.14 and 0.12 in GDH and LADH respectively - consequently cavities (or rather lack of) do not represent a feature that confers stability in GDH.

However, the intersubunit cavity present in the GDH dimer interface is quite interesting (and not present in LADH). The cavity is considerable ($\approx 64 \text{\AA}^3$), and is situated within the intersubunit aromatic/aromatic cluster. Analysis of the solvent accessibility of the residues lining this cavity suggest that the cavity is empty (that is not water-filled). Thus despite the fact that GDH is thermostable, and has a greater degree of hydrophobic dimeric interactions, the presence of this cavity must be a destabilising influence. This represents a prime target for future cavity filling mutations to attempt to further increase the thermostability of GDH (see future studies).

Putative Archaeal features - the minimal functional unit.

To be able to discern possible features that may be unique to the archaeal enzymes, at least two structures of archaeal enzymes need to be solved. Recently, two structures

of archaeal enzymes have been solved, GDH and citrate synthase (CS) (Russell *et al.*, 1994) from *Thermoplasma acidophilum*.

The emergence of these two structures have revealed that the enzymes possess similar features that may represent an archaeal characteristic, and have led to the proposal of *the minimal functional unit*, which is defined as the structural framework required to maintain the structural integrity and retain the catalytic activity for a given enzyme.

The structures of these archaeal enzymes possess similar features that may represent an Archaeal characteristic. CS displays a dramatic reduction in the size of the loops connecting the α -helices, resulting in a more compact and stable enzyme, whilst maintaining similar active site conformations. The nucleotide-binding domain of GDH (see chapter 7) consists of fewer secondary structural elements/ smaller loops that were deemed necessary for binding the coenzyme, NAD(P), in comparison to previously reported domains.

The Archaea are an evolutionary primitive domain of organisms that generally live in harsh environments, and so the evolutionary pressure to maintain the structural integrity of an enzyme would be great - any mutations that would, for example, reduce the thermostability of an enzyme, would not be permitted. Evolution of organisms into less harsh environments would have allowed enzymes the ability to accommodate adaptations to the minimal functional unit that may provide additional regulation of the existing enzymic activity.

Structures of more archaeal enzymes are required to confirm the proposal of *the minimal functional unit* (Russell & John, 1994).

Chapter 10

BIOCHEMICAL STUDIES

Introduction

The main purpose of this chapter is not to provide comprehensive kinetic, fluorescence and thermostability studies, but to gain evidence that GDH is biochemically (as well as structurally) related to LADH.

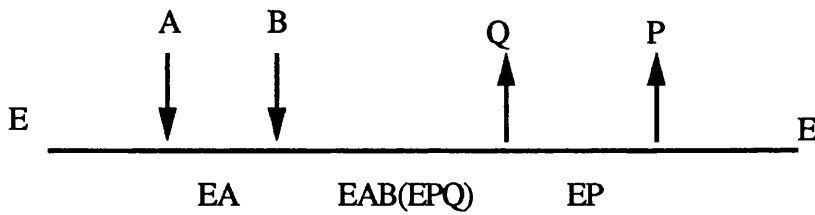
Determination of zinc

The presence of zinc within GDH was determined qualitatively and quantitatively using the Oxford Microprobe and flame absorption atomic spectroscopy (appendix 1, zinc determination) respectively. The latter method has been used extensively to accurately quantify the metal content within proteins (Vallee & Auld, 1990; Falchuk *et al.*, 1988).

A value of 7.76 ± 0.19 zinc ions/tetramer was determined, which based on the LADH structure, strongly suggests that the GDH monomer contains 2 zinc ions: a structural and catalytic zinc.

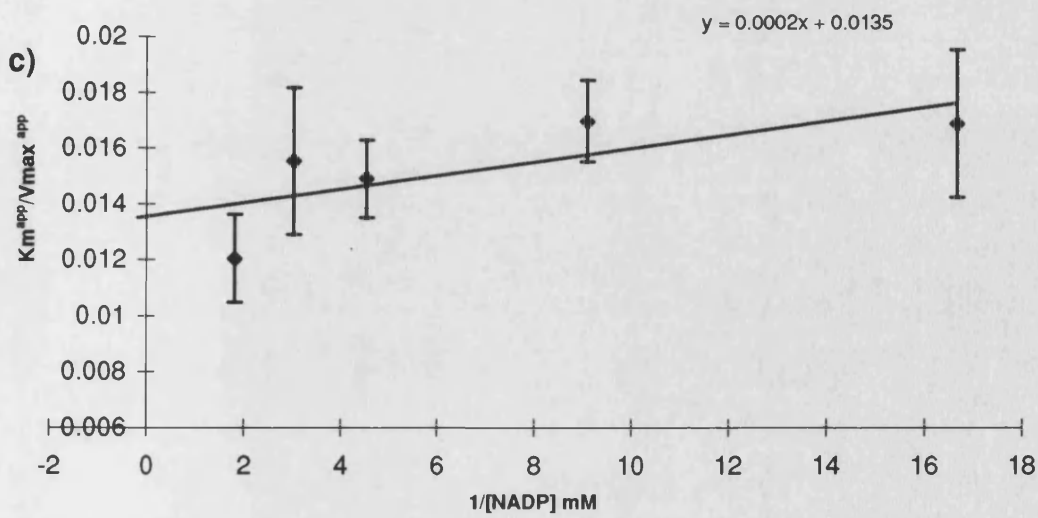
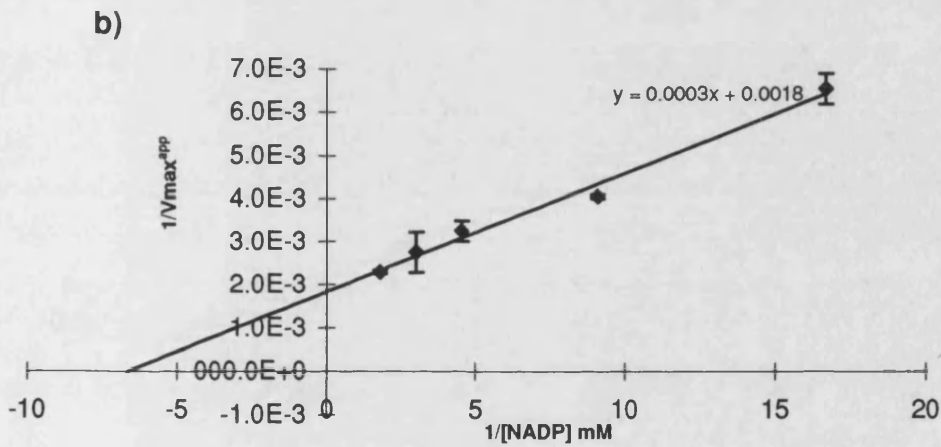
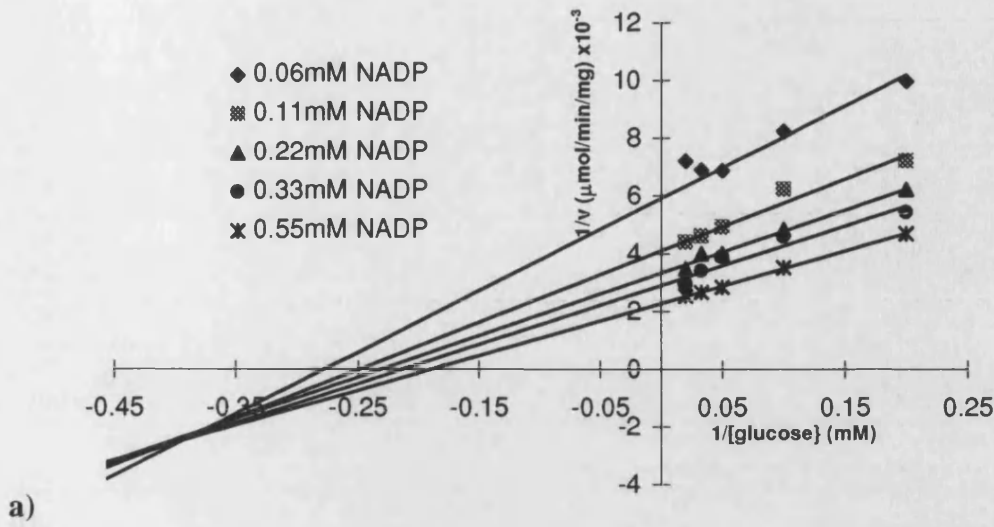
Determination of mechanism

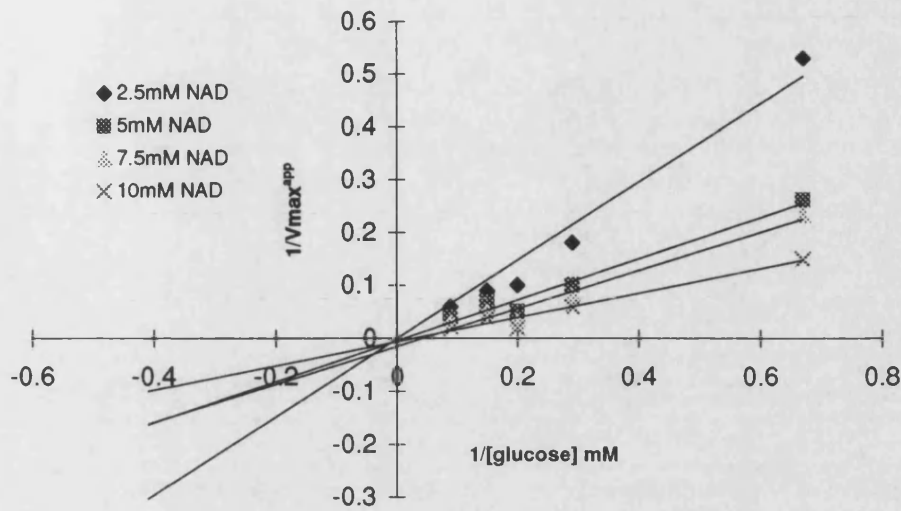
The Cleland notation for a compulsory order mechanism is:



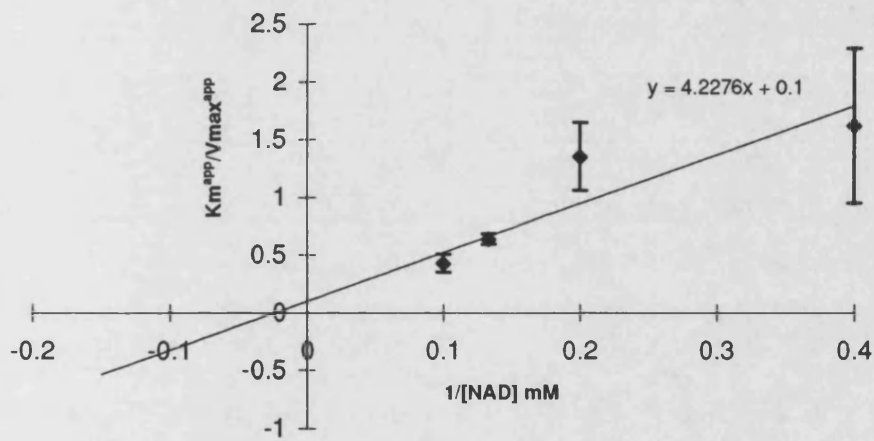
LADH proceeds via a slight variation of the compulsory order mechanism (the Theorell-Chance mechanism (1951) in which the interconversion and breakdown of the ternary complexes are so rapid that EAB and EPQ are kinetically undetectable (A, NAD; B, alcohol; Q, aldehyde; P, NADH).

It was postulated that the GDH catalysed reaction also proceeds via the compulsory order mechanism (A, NAD(P); B, glucose; Q, gluconolactone; P, NAD(P)H). The mechanism of GDH was determined kinetically (according to the method of Henderson, 1992) (appendix 1, Biochemical methods). The initial primary and secondary kinetic plots indicated that the mechanism was either compulsory order or random equilibrium (Wharton & Eisenthal, 1981). To distinguish between these two mechanisms, and to discern the order of substrate binding, product inhibition studies using gluconolactone (gluconic acid did not inhibit the enzyme) were undertaken. The product inhibition plots were only used as a diagnostic to discern the mechanism (no kinetic constants were determined). Plot 10.1g is indicative of uncompetitive inhibition, whilst plot 10.1h) is diagnostic of simple non-competitive inhibition (Henderson, 1992) - which suggests that GDH proceeds via the compulsory order mechanism, in which NADP binds first, and glucose second. Further kinetic studies are required to determine whether GDH proceeds via the Theorell-Chance mechanism as opposed to the 'standard' compulsory order mechanism.

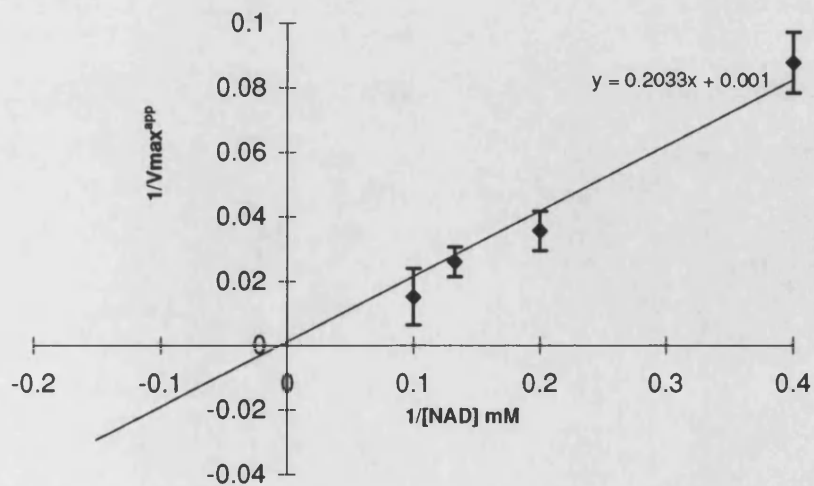




d)



e)



f)

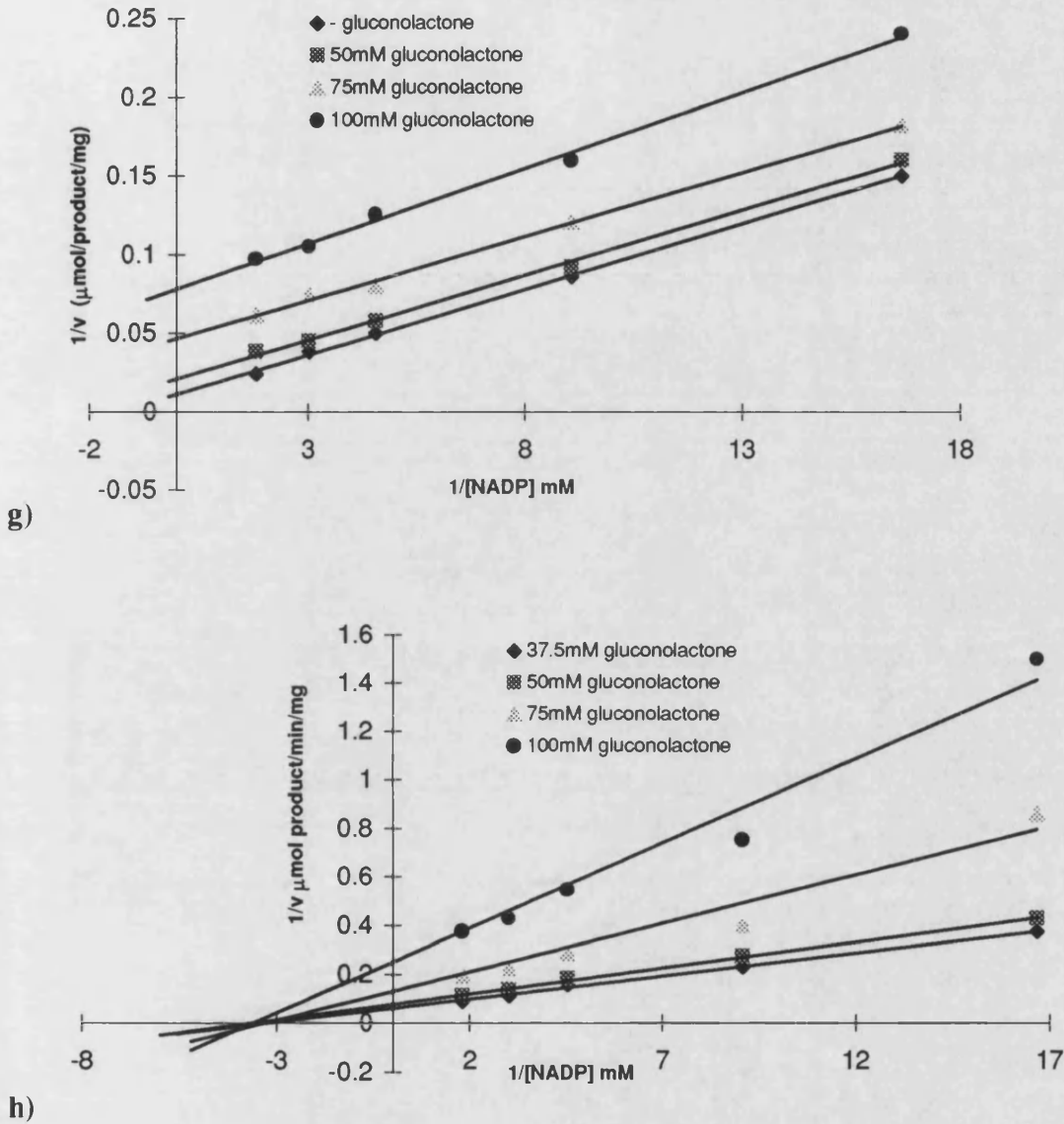


Fig 10.1 (a to h) Kinetic plots used to determine the kinetic constants and the mechanism of the GDH catalysed reaction. a) and d) are the primary double reciprocal plots for NADP and NAD respectively. These primary plots were used for display purposes only - the values of the K_m^{app} and V_{max}^{app} were determined from the direct linear plots. Replotting of these *apparent* kinetic constants (the **secondary plots**) b) and c) for NADP, e) and f) for NAD yields the true kinetic constants. Plots g) and h) are the primary double reciprocal plots of the product inhibition studies (saturating and unsaturating glucose respectively), which were diagnostic of the compulsory order mechanism.

The steady-state kinetic constants

The K_m values for NADP and glucose, as well as the V_{max} values of GDH were calculated from the secondary plots, and were in accordance with the previously reported constants (table 10.1). In addition, the K_m for NAD, which had not been previously accurately determined, was calculated to be 20.3mM.

GDH has an apparent low affinity for the coenzyme in comparison to other dehydrogenases - with the exception of YADH, which possessed similar kinetic constants (Tsai *et al.*, 1989). Despite LADH displaying a greater affinity for the cofactor than GDH and YADH, LADH possesses a much lower specific activity.

Enzyme	$K_m^{NAD(P)}$ (mM)	$K_m^{substrate}$ (mM)	V_{max} ($\mu\text{mol product/min/mg}$)
GDH [*]	0.17/20.3	7.51	555
GDH ^a	0.29	9.40	457
GDH ^b	0.11	10.30	320
LADH	6.26×10^{-3}	0.44	139
YADH	0.228	2.25	300

Table 10.1 Summary of the kinetic constants of GDH (NADP/NAD), LADH (NAD) and YADH (NAD). * denotes kinetic constants determined in this study, ^a by Bright *et al.*, (1993); ^b by Smith *et al.*, (1989).

Thus the tight binding appears to be detrimental towards activity (Klinman, 1981).

GDH, YADH and LADH appear to possess a similar number of interactions to the coenzyme, therefore there must be another factor affecting coenzyme binding.

This functional difference may arise from the differing electrostatic potentials exhibited by GDH and LADH (chapter 8, GRASP). From the surface potential plots, it was observed that the coenzyme binding sites of LADH and GDH were positively and

negatively charged respectively. The former would serve to attract the NAD, whereas the negatively charged potential of GDH would repel the incoming NADP.

This GDH feature appears to be a biological anomaly - however it may serve to increase the V_{\max} of the reaction.

The difference between NAD(P) and NAD(P)H is a hydride, and consequently the enzymes' affinity for each coenzyme form cannot differ enormously, which would explain why the dissociation of the reduced coenzyme from LADH is rate limiting (Theorell & Chance, 1951). Whereas in GDH, the repulsive coenzyme-enzyme forces result in GDH possessing a low affinity for NADP, which would ensure a rapid removal of the NADPH. Whether YADH exhibits similar electrostatic features to GDH awaits to be seen.

The conformational change

LADH undergoes a conformational change upon binding of NAD (Theorell & Tatemoto, 1971) triggered by correct positioning of the nicotinamide ring (Eklund & Branden, 1979, and references therein). Tryptophan fluorescence studies have been used to investigate this conformational change in GDH. Fig 10.2 shows that a saturating concentration of NADP resulted in a 60% quenching of the tryptophan fluorescence in GDH, and furthermore this fluorescence quenching is dependent on the concentration of NADP.

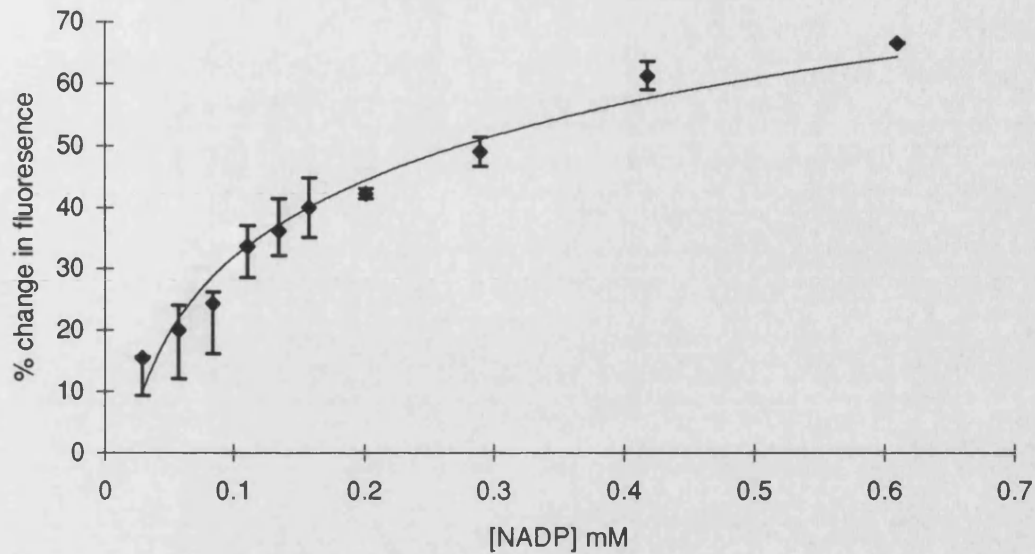


Fig. 10.2 The effect of NADP on the fluorescence of GDH. An NADP dependent fluorescence quenching was observed.

GDH possesses one tryptophan per monomer - Trp 321, situated in the interconnecting helix of the C-terminal catalytic domain, which is distant from the active site, and buried in a hydrophobic environment. Any change in the tryptophan fluorescence, promoted by the binding of NADP to GDH, must therefore be a consequence of conformational change, in which the tryptophan is transferred to a more solvent exposed environment.

A comparison between the apo-form and the modelled closed form of GDH reveals that the tryptophan environment has altered, which may explain the observed tryptophan fluorescence quenching (fig. 10.3).

Pre-steady state fluorescence and kinetic studies were undertaken in addition to the steady state studies above (initially to ascertain whether GDH would be a good model system for Dye-binding experiments).

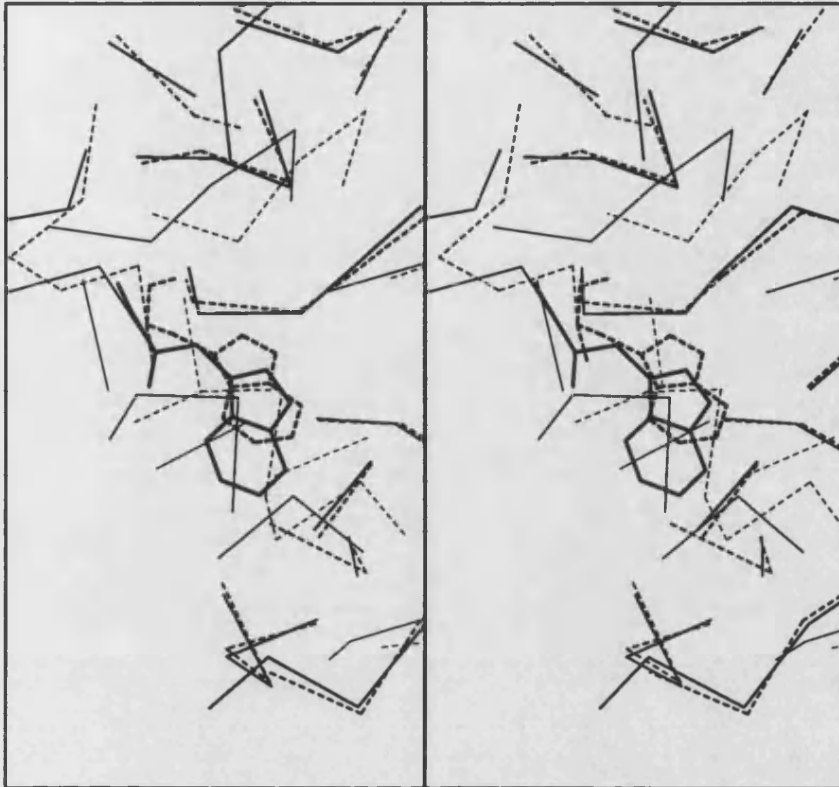


Fig. 10.3 Putative movement of the tryptophan residue between the open (solid lines) and closed form (dashed lines).

In LADH, the coenzyme induced conformational change is slow, with the rate of conformational change having been studied by pre-steady state tryptophan fluorescence studies (Theorell & Tatemoto, 1971). This rate of conformational change could not be measured for GDH (using pre-steady state tryptophan fluorescence studies), because the conformational change was too rapid - only the steady state fluorescence quenching was observed, as opposed to a pre-steady state burst monitoring the conformational change. Significantly, the rate of conformational change was observed in LADH (data not shown).

Analysis of the pre-steady state kinetics of GDH revealed a significant lag-phase in the GDH catalysed NADPH production.

This NADP-induced conformational change in GDH supports the kinetic data in which NADP binds first in the compulsory order mechanism. In addition, this rapid conformational change may indicate that the rate-limiting step in GDH is not the conformational change - suggesting that the binding of glucose, or the rate of hydride transfer is the rate-limiting step.

A structurally interesting observation is that NADP is surrounded by a number of aromatic residues in GDH(fig. 10.4)(which are not present in LADH) - whether this has any consequences for the rapid conformational change requires further study.

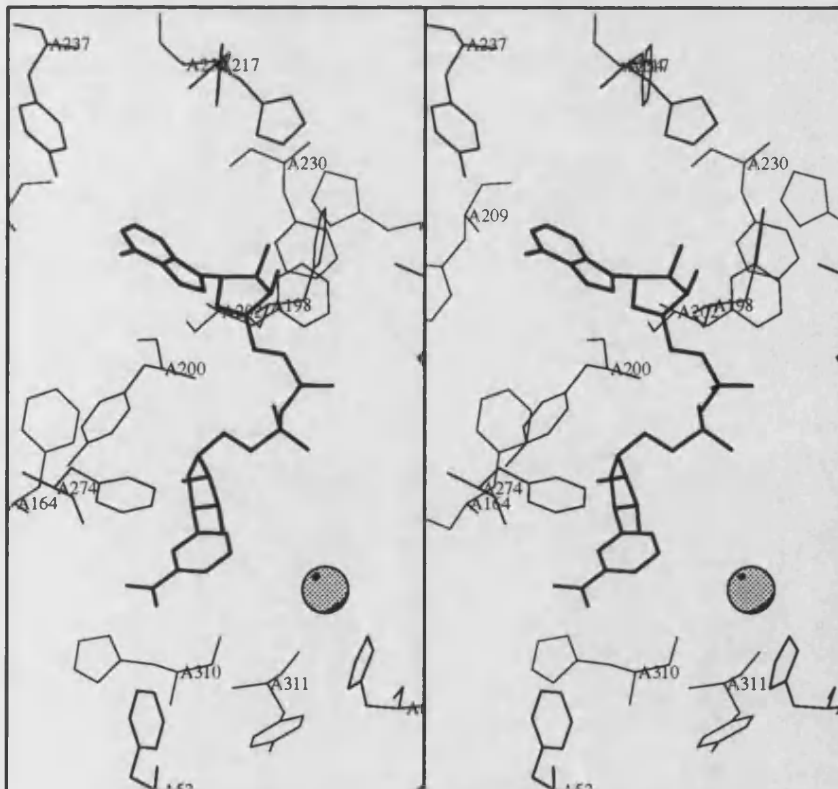


Fig. 10.4 The aromatic/NAD interactions.

GDH has more similarity to YADH than LADH

Interestingly, the rate limiting step in the YADH mechanism is the rate of hydride transfer, and not the coenzyme-induced conformational change (Klinman, 1981). It is possible that the hydride transfer step is the rate limiting step in the GDH reaction. This may be rationalised by the features present in the GDH active site (Chapter 8) : GDH does not possess a residue equivalent to the proton relay His 51 found in LADH (Ehrig *et al.*, 1991). In addition, in the substrate modelling, the C1-hydroxyl of glucose did not directly bind to the catalytic zinc - although there is the possibility of an indirect interaction via a water molecule. The substrate of LADH was shown crystallographically to directly bind to the catalytic zinc, however other studies have pointed towards a water molecule participating in the interaction (Eklund *et al.*, 1982, and references therein). Further detailed studies are necessary (future work).

The similarity in steady state kinetic constants, the rapidity of the conformational change, and the rate-limiting hydride transfer step suggests that GDH is mechanistically more similar to YADH than LADH.

The structural and phylogenetic similarities also fortifies the mechanistic implications - that GDH is more homologous to YADH (and SDH) than LADH. Further studies are necessary to confirm this hypothesis (see future work).

Thermostability studies

The structural zinc in GDH appears to have a defined role in maintaining the quaternary structure. A study has shown that the structural zinc of the tetrameric YADH plays an important conformational role (Magonet *et al.*, 1992). Magonet *et al.*

have shown that the selective removal of the structural zinc is possible (using dithiothreitol (DTT)), which results in a time and concentration dependent concomitant reduction in the thermostability of the enzyme. The following series of experiments attempted to mimic the studies of Magonet *et al.*, and ascertain whether the structural zinc effects the conformational stability of GDH.

The chelating agent

Magonet *et al.* selectively removed the structural zinc using the chelating agents DTT and EDTA. The latter removes the structural zinc only at low temperatures (4°C), where the catalytic zinc is inaccessible to the EDTA. Higher temperatures (30°C) removes both the catalytic and structural zinc.

As GDH is thermostable, it was believed that at 30°C, the enzyme would be too inflexible (rendering the catalytic zinc inaccessible to the EDTA) to permit removal of both the zinc ions by the EDTA.

Stability assays

[EDTA]mM	0	10	20	30	50	100
30°C	104	93	98	91	88	77
55°C	100	68	56	50	44	27

Table 10.2 Effect of % GDH activity after a 1 hr incubation of varying concentrations of EDTA concentrations at 30°C and 55°C.

GDH activity is largely unaffected by concentrations EDTA < 30mM at 30°C, whereas high concentrations of EDTA cause an $\approx 25\%$ reduction of activity at this temperature (table 10.2). At 55°C, there was a marked reduction in GDH activity, with an

approximately 50% reduction of activity after incubation with 20-30mM EDTA. Two conclusions can be drawn from this:

- At 55°C, 20mM EDTA can now chelate the catalytic zinc after a 1hr incubation, whereas at 30°C it could not.
- At 30°C, the 20mM EDTA chelates only the more solvent exposed structural zinc (which would not affect the enzymic activity). However, removal of this structural zinc reduces the thermal stability of the enzyme.

Time (min)	0	10	20	30	40	50	60	85
- EDTA	99	96	107	100	103	101	105	96
+ EDTA	100	75	72	68	64	66	58	56

Table 10.3 % decrease in GDH activity at various time intervals in the absence and presence of 20mM EDTA at 55°C.

Table 10.3 shows the time-dependent effect of 20mM EDTA on GDH activity at 55°C. GDH was shown to be thermostable at 55°C in the absence of EDTA, whereas in the presence of EDTA, a sharp fall in activity was observed after 10 min, followed by a more gradual loss of activity. A 10 min incubation with 20 mM EDTA was believed to represent the optimal incubation time for selectively removing the structural zinc.

This sharp fall in activity suggests that the selective removal of the structural zinc affects the thermostability of GDH.

In a further experiment, GDH was incubated at 30°C with 20mM EDTA for 1hr, and then heat-shocked for 10 minutes (table 10.4).

Temp°C	30	40	50	60	70	80
- EDTA	100	98	95	95	67	58
+ EDTA	95	92	87	63	40	7

Table 10.4 Effect of heat-shocking the EDTA treated GDH sample.

GDH, in the absence of EDTA was stable up to 60°C, but began to lose activity at higher temperatures. In the EDTA treated sample, GDH was noticeably less thermostable - nearly losing all activity at 80°C, indicative of a marked difference between the EDTA treated and untreated GDH samples.

These results are best explained by the selective removal of the structural zinc, which does not effect the catalytic zinc (and hence activity) of the enzyme, but renders the enzyme less thermotolerant. Removal of the structural zinc could easily be perceived to destabilise the structural lobe, which would in turn destabilise the quaternary structure of GDH. Further experiments (see future work) are necessary to unambiguously discern the structural role played by the structural zinc.

Appendix 1

Protein methods

Production of cell free extracts

TG1 E.coli cells containing the glucose dehydrogenase gene in the pMex 8 plasmid were harvested from an overnight culture by centrifugation (5000rpm/ 10min) and resuspended in 50mM sodium phosphate buffer pH7. The cells were lysed by sonication on ice for 1 minute(4x15 sec) with the resultant debris removed by centrifugation (12000g /30min).

GDH assay

GDH activity was determined spectrophotometrically by monitoring the production of NADPH at 340nm, at 55°C, with respect to time. The standard 1ml assay mixture contained: 0.5mM NADP⁺ , 50mM D-glucose, 50mM sodium dihydrogen orthophosphate buffer, pH 7.0.

Protein estimation

Protein concentration was determined spectrophotometrically using the method of Bradford at 595nm. 0 to 20µg solutions of Bovine serum albumin were used as the standard curve.

For a more accurate estimation of protein concentration, the U.V. absorbance of the protein at 280 and 260nm was determined.

Heat treatment

The cell free extract was heated at 65°C for 10 min in the presence of 50mM D-glucose. The sample was incubated on ice for 10min and centrifuged (10000g/ 20min). The pellet was discarded.

Solvent denaturation

Ice cold methanol was added to the sample to a final concentration of 50 %(v/v) methanol. The sample was incubated on ice for 10min and the precipitated E.coli proteins removed by centrifugation

(10000g/20min). The recombinant glucose dehydrogenase was precipitated in 80%(v/v) methanol and recovered by resuspension of the pellet in 50mM sodium phosphate pH7. Any undissolved material was removed by centrifugation (10000g/10min).

Gel filtration

The sample (4ml total volume) was loaded onto the FPLC Hiload 16/60 Superdex 200 gel filtration at a flow rate of 1ml/min. The glucose dehydrogenase was eluted as a single peak of activity.

Anion exchange

The sample was loaded onto the FPLC MonoQ column at a flow rate of 1ml/min. The loaded column was washed with 30ml of 50mM sodium phosphate buffer pH7. The bound protein was eluted using a 30ml 0-0.2M sodium chloride gradient in the same buffer; any remaining bound protein was eluted with 10ml of 2M NaCl.

Dye ligand chromatography

The sample was loaded onto the Matrex red gel A (10cm) at a flow rate of 1ml/min. The loaded column was washed with 40ml of 50mM sodium phosphate buffer pH 7. The bound protein was eluted using 7.5ml of 15mM NADP.

SDS-polyacrylamide Gel Electrophoresis

Discontinuous SDS-polyacrylamide gel electrophoresis was performed according to the method of Laemmli (1970). The gels were fixed and stained , then left to destain overnight.

Crystallisations

Hanging drop

The samples were concentrated to 6-20mg/ml protein using the centricon concentrators prior to crystallisation trials. The hanging drop vapour diffusion method using Limbro tissue culture plates

(McPherson, 1990) was employed in setting up the crystallisation trials. The cover slips had to be silanised prior to the crystallisation trials.

Macroseeding

The macroseeding technique is best displayed as a diagram (Fig. 1)

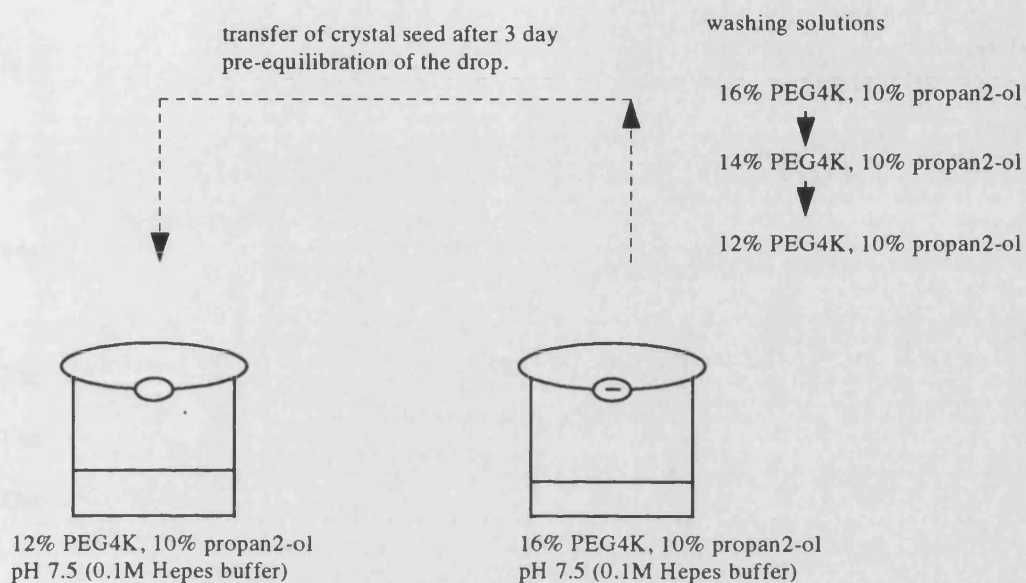


Fig. 1.1 The macroseeding technique (not to scale).

Small crystals of GDH were grown in 16% PEG, 10% propan-2-ol, pH 7.5. To establish the optimum conditions for seeding, varying concentrations of precipitant (16 to 8% PEG) were equilibrated against a protein/precipitant drop for 0, 1, 2, 3, 5 days. A crystal seed was then transferred to each pre-equilibrated drop. The optimum conditions for transferring the crystal seed were found to be when a protein had been pre-equilibrated against 12% PEG, 10% propan-2-ol for 3 days

Biochemical methods

Primary plots

Initial steady-state rates were determined at glucose and NAD(P) concentrations ranging from $K_m/2$ to $8K_m$. Range of glucose, NADP and NAD concentrations were 5 to 50mM, 0.06 to 0.55mM, 2.5 to 10mM

respectively. The steady state rates were determined as the [glucose] was varied at each fixed concentration of cofactor .

Assuming that the reaction proceeded via a compulsory order mechanism, at a fixed concentration of NAD(P) in the absence of product, the rate equation is:

$$v = \frac{V_{\max}[\text{glucose}][\text{NAD(P)}]}{K_m^{\text{glucose}}[\text{NAD(P)}] + K_m^{\text{NAD(P)}}[\text{glucose}] + [\text{glucose}][\text{NAD(P)}] + K_{ia}K_m^{\text{glucose}}}$$

In reciprocal form the rate equation is:

$$\frac{1}{v} = \left[\frac{K_m^{\text{glucose}}}{V_{\max}} \left(1 + \frac{K_{ia}K_m^{\text{NAD(P)}}}{K_m^{\text{glucose}}[\text{NAD(P)}]} \right) \right] \frac{1}{[\text{glucose}]} + \frac{1}{V_{\max}} \left(1 + \frac{K_m^{\text{NAD(P)}}}{[\text{NAD(P)}]} \right)$$

where slope = $K_m^{\text{app}}/V_{\max}^{\text{app}}$ and intercept = $1/V_{\max}^{\text{app}}$

The values of the kinetic constants in the above equation are apparent values only (V_{\max}^{app} and K_m^{app})

The V_{\max}^{app} and K_m^{app} values were calculated from the Direct linear plot using the ENZPACK software.

The primary double reciprocal plot was used for display purposes only.

Secondary plots

From the above equation, it can be seen that plotting $1/V_{\max}^{\text{app}}$ and $K_m^{\text{app}}/V_{\max}^{\text{app}}$ against the reciprocal [NAD(P)] would yield the true kinetic constants.

$$\frac{1}{V_{\max}^{\text{app}}} = \left(\frac{K_m^{\text{NAD(P)}}}{V_{\max}} \times \frac{1}{[\text{NAD(P)}]} \right) + \frac{1}{V_{\max}}$$

$$\frac{K_m^{\text{app}}}{V_{\max}^{\text{app}}} = \left(\frac{K_{ia}K_m^{\text{NAD(P)}}}{V_{\max}} \times \frac{1}{[\text{NAD(P)}]} \right) + \frac{K_m^{\text{glucose}}}{V_{\max}}$$

Product inhibition

2 product inhibition studies were performed at (i) a saturating glucose concentration (50mM) and (ii) an unsaturating glucose concentration (3.3mM) over a range of NADP concentrations (0.06 to 0.55mM).

The gluconolactone was prepared fresh and kept on ice prior to the assay. Gluconolactone assay concentrations of 37.5, 50, 75 and 100mM were used in the experiment. The primary double reciprocal plots were used as a diagnostic to discern the mechanism only.

Steady state fluorescence studies

Fluorescence studies were performed on a spectrofluorimeter with the assistance of R. Maytum. The fluorescence level of a 0.03mg/ml GDH sample was measured at a range of NADP concentrations (0 to 0.6mM), at 25°C, in 50mM sodium dihydrogen orthophosphate buffer. Excitation wavelength at 280nm, emission wavelength at 330nm.

Pre-steady state kinetic and fluorescence studies

These experiments were performed by R. Maytum. The conditions used in the steady state kinetic and fluorescence studies were mimicked. The author confesses no detailed knowledge of the functioning of this apparatus.

Zinc determination

Proton microprobe (nuclear physics lab, Oxford)

Analysis performed by E. Garman. A GDH crystal was bombarded by a high energy proton beam. The characteristic X-rays emitted are used for elemental analysis of the sample. Reservoir buffer was used as the control.

Flame atomic absorption spectrometry

Performed by a Dr. A. Carver. 1-2ml of the GDH sample was aspirated into the flame-nebulizer system of the spectrometer, where the zinc atoms are atomized. The zinc atoms absorb the monochromatic radiation (213.9nm for zinc). The amount of absorbance is proportional to the [zinc]. Standard zinc solutions were prepared by A. Carver.

Thermostability studies

Varying EDTA concentration

The GDH sample was incubated for one hour in the presence of a range of EDTA concentrations (0 to 100mM) at 30°C and 55°C . The residual GDH activity was then measured.

Time-dependent effect of EDTA

GDH sample was incubated from 0 to 85 minutes in the presence and absence of 20mM EDTA at 55°C. Residual GDH activity was then measured.

Heat-shock with EDTA

The GDH sample was incubated in the presence and absence of 20mM EDTA for one hour. The sample was then heat shocked for 10 minutes (temperature ranging from 30 to 80°C). Residual GDH activity was then measured.

Phylogenetic tree construction

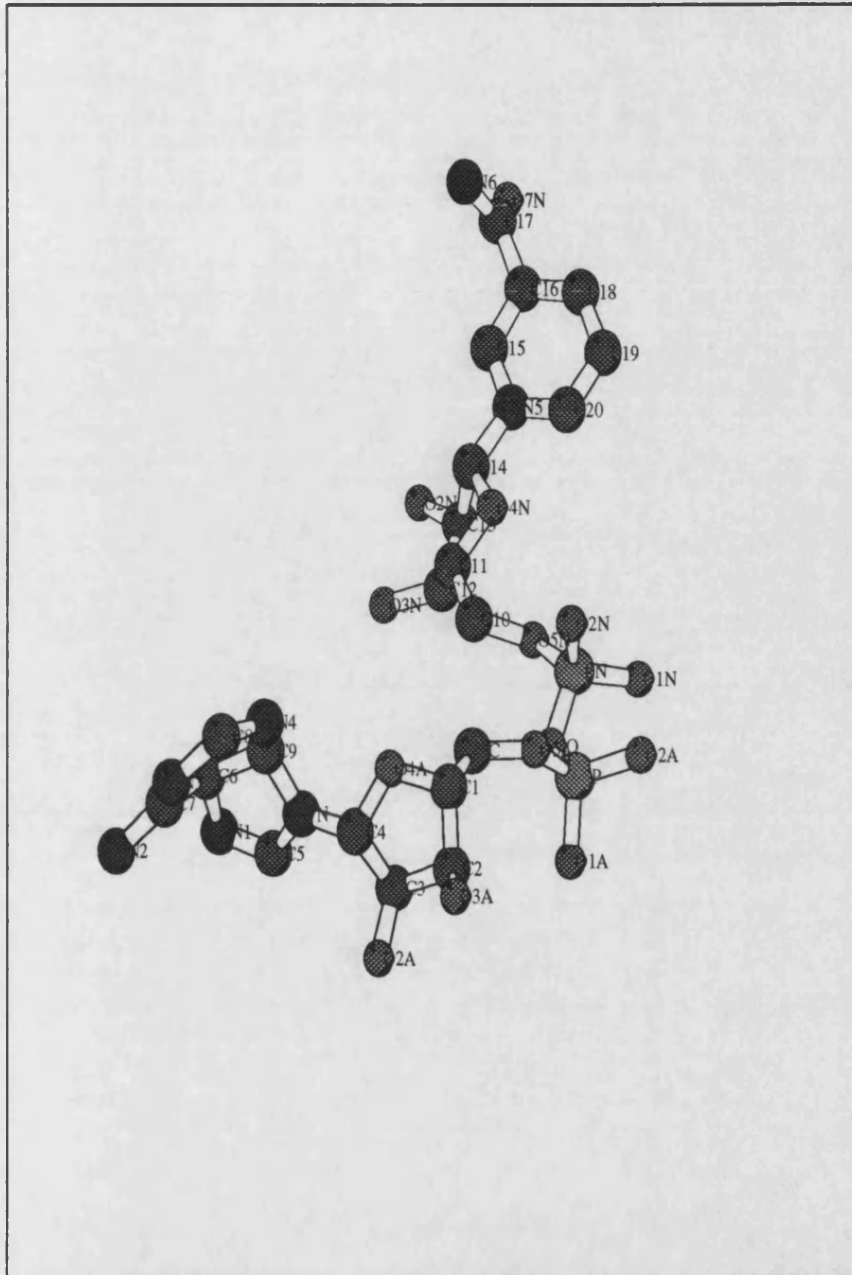
The tree was derived from the structurally based alignment using the **PHYLIP** package of programs. Protein distance matrices were first calculated (**PROTDIST**), from which **FITCH** analysed these distances and computed the trees. **DRAWTREE** displayed the unrooted tree.

Molecular modelling

Minimisation

The minimisation was carried out using **INSIGHT** with the aid of P. Calleja. The minimisation was carried out in stages. First, the protein heavy atoms were tethered to their experimental positions by a penalty function - which allowed relaxation of the modelled glucose and NAD(P) without disturbing the overall protein structure. The all constraints were removed and minimisation was performed using a steepest descent algorithm to remove any remaining large strain. Finally unconstrained minimisations were performed using a conjugate gradient method until convergence was achieved.

Substrate/NAD structures



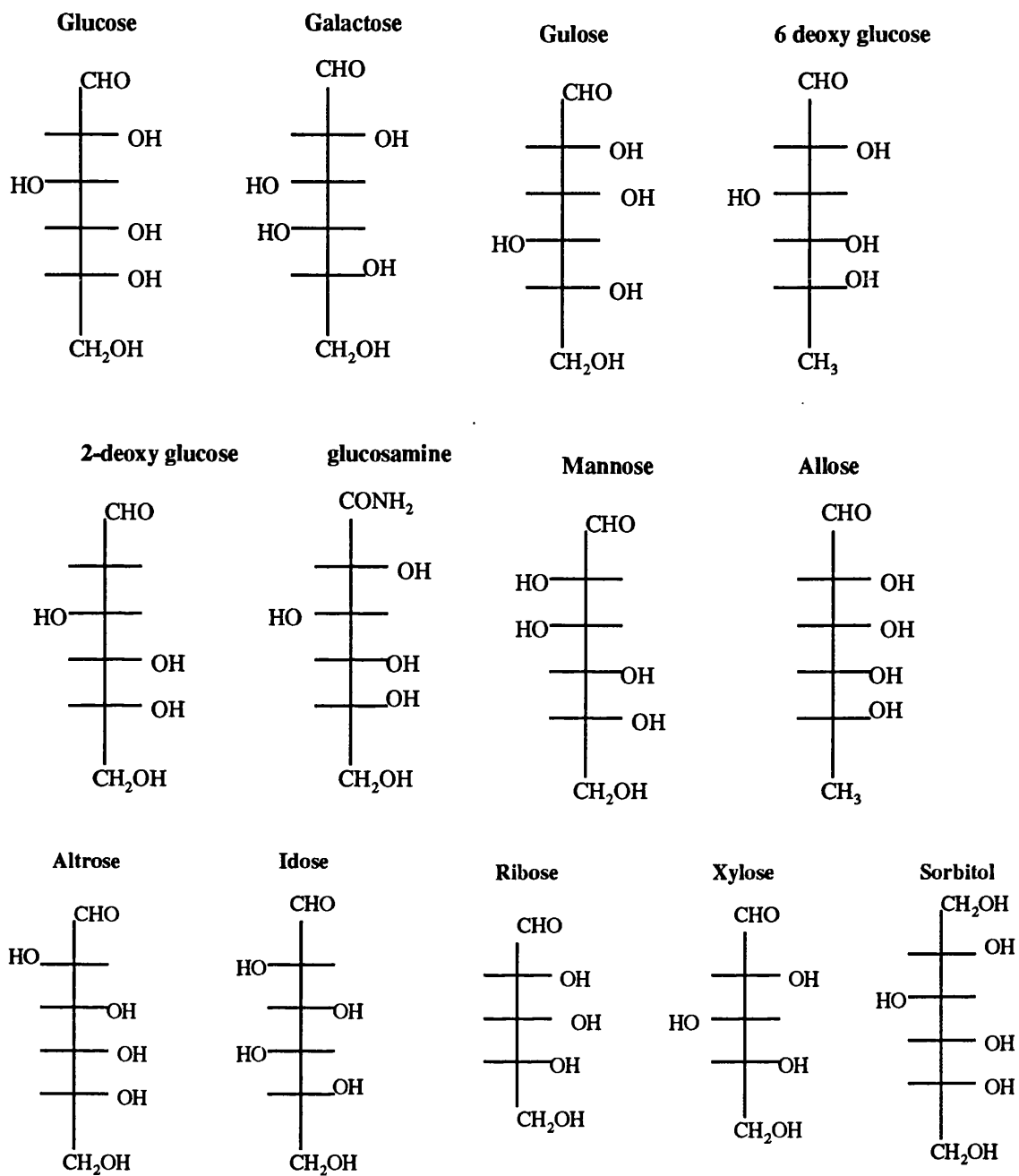


Fig. 1.2 Structures of NAD and the sugars used in the specificity studies.

Programs/software used

Crystallisations

INFAC - a program that establishes a crystallisation matrix (donated by W. Carter). Analysis of the output was performed by W. Carter.

Data processing and analysis

XDS - Kabsch's program that processes data collected on the area detector.

DENZO - Otwinowski's package that processes data collected on the image plate (and film).

XSCALE - Kabsch's program that merges data sets

POLARRFN* - Kabsch fast polar rotation function.

Derivative processing

X2L* - converts hkl files to LCF format.

LCF2MTZ* - converts LCF format to MTZ format

MTZUTILS* - a program that combines 2 reflection files, or edit columns, labels or titles.

MTZDUMP* - lists header and reflections to the terminal.

MTZTONA4* - converts MTZ files to portable NA4 ASCII format.

CAD* - Combine assorted data (and sort) a number of reflection files . It was used to convert data from one area of reciprocal space to another.

LOCAL - local scaling of 2 data sets.

FFT* - computes the crystallographic fast Fourier transformation.

PLUTO* - used to plot defined sections of the Patterson maps.

VECSUM* - an automated Patterson peak search.

VECREF* - Patterson space refinement of the heavy atom sites.

MLPHARE* - Refines the heavy atom parameters and calculates phases.

FHSCAL* - scales native to derivative data.

SIGMAA* - phase combination of isomorphous and calculated phases.

PEAKMAX* - searches for peaks in the electron density map

Map and structure factor calculation

SFALL* - structure factor calculation from map or atomic co-ordinates using inverse FFT

RSTATS* - least squares scaling between F_o and F_c .

EXTEND* - extends the asymmetric unit to cover any grid volume

RESEC - resections the map

RUNMAPAGE - converts the maps to brick (O) format.

Phase improvement

SQUASH - phase improvement and extension by various methods.

WANG - Wang's suite of programs for solvent flattening.

Molecular averaging

RAVE - suite of programs from Uppsala for non-crystallographic averaging.

MAMA - suite of programs from Uppsala for mask creation and manipulation.

MAPMAN - program from Uppsala for manipulation of electron density files.

Model building

O - a graphics package used in conjunction with an Evans + Sutherland PS390 ESV-10 for the visualisation and interpretation of the electron density maps.

BONES - computes a skeletonised representation of the electron density.

Refinement

X-PLOR (version 2.1) - Brünger's simulated annealing refinement package :

GENERATE - generates hydrogens and a topology file.

CHECK - determines a weighting factor.

PREPSTAGE - conventional least squares minimisation.

SLOWCOOL - simulated annealing refinement.

Assessment of model quality

PROCHECK* - a comprehensive package that checks the stereochemistry of the model.

LUZZATTI - program within X-PLOR that generates a Luzzatti plot.

3D-1D PROFILE - determines whether the model makes biochemical sense.

RS_FIT - program within O that calculates model/map correlation.

Miscellaneous

MOLSCRIPT - a program to produce pretty pictures.

RASTER 3-D - a program to produce very pretty pictures.

DSSP - determines the secondary structure of the protein from atomic co-ordinates.

HERA - produces a topology file from the DSSP output.

CONTACT* - calculates various types of contacts in a structure.

ACCESS - calculates the solvent accessibility of residues within a structure.

SHP - structural superposition of two structures.

GRASP - a software package that enables a graphical representation and analysis of surface properties.

GCG -University of Wisconsin Genetics Computer Group: a sequence analysis software package

AMPS - Geoff Barton's multiple sequence alignment package

ALSCRIPT - produces attractive displays of multiple sequence alignments (Geoff Barton).

COMPOSER - Birkbeck molecular modelling software. In this project it was used in multiple structural superposition of the nucleotide binding domains.

Garry's jiffys - numerous FORTRAN programs devised by Garry that generally made computing life easier.

* denotes programs from the CCP4 (Collaborative Computing Project Number 4) suite.

This thesis was created using the Microsoft Office TM package on a *Genie professional* 4DX266 PC.

Word-processing using Microsoft Word (6.0), Flowcharts using Powerpoint (4.0), Graphs using Excel (5.0)

Hardware

Microvax 4000, Microvax 3300, Silicon Graphics Indigo and an Evans + Sutherland PS390 ESV-10.

Materials

General reagents and salts for growth media, unless otherwise specified, were from BDH,, SIGMA., or FISONS, U.K.; yeast extract and tryptone were from Difco, Michigan U.S.A. NAD(P) (grade 1) were from Boehringer.

Appendix 2

The unit cell

The basic building block of a crystal is termed the unit cell, which is the smallest element that is completely representative of the whole crystal. The array of points at the vertices of the unit cell is called the crystalline lattice. a , b , c define the lengths of the unit cell and α , β , γ define the unique angles. There are seven different types of unit cell, triclinic, monoclinic, orthorhombic, tetragonal, cubic, trigonal and hexagonal. GDH crystals were monoclinic, where $a \neq b \neq c$, $\alpha = \gamma = 90^\circ$, $\beta > 90^\circ$. The unit cell of a crystal is in real space, whereas the diffraction pattern relating to the unit cell is in reciprocal space. The reciprocal space lattice (a^* , b^* , c^*) is related to the real space lattice - for a cell where all angles are 90° , $a^* = 1/a$, $b^* = 1/b$, $c^* = 1/c$. A large unit cell results in a small reciprocal cell, with the reflections being closely spaced in the diffraction pattern.

In real space, the indices hkl define a particular set of equivalent, parallel planes, and specifies the number of planes that exist per unit cell in the a , b , c directions respectively. The Miller indices hkl are used to designate the individual reflections in reciprocal space of the diffraction pattern.

Symmetry within the GDH unit cell

GDH belongs to the monoclinic space group, $P2_1$, which means that an entity (the GDH tetramer) within the unit cell is related to another tetramer in the unit cell by a 2-fold screw axis (which is along b). This can be represented by the symmetry operators (x, y, z) and the symmetry equivalent $(-x, y + 1/2, -z)$ (fig. 2.1). This real space symmetry is reflected by symmetry in the diffraction pattern (reciprocal space). For $P2_1$, this also results in $0k0$ reflections being systematically absent when k is odd.

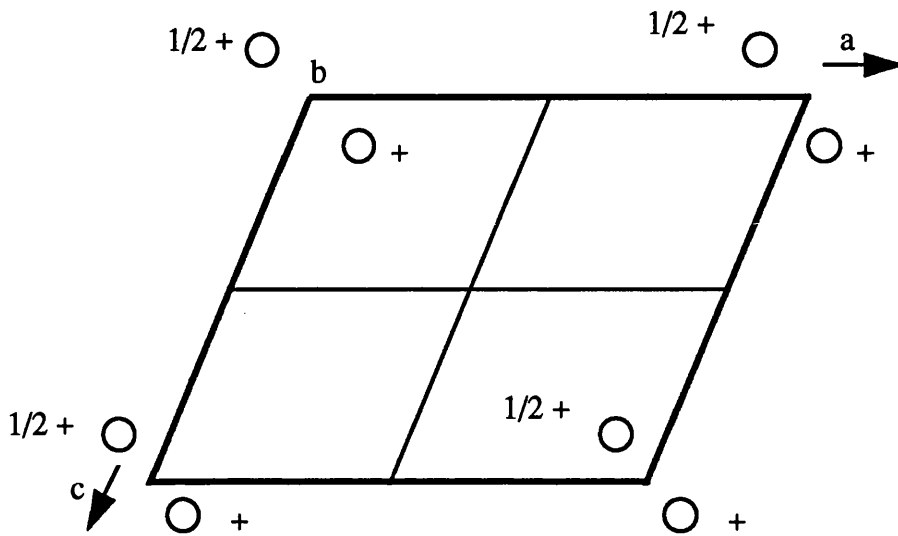


Fig. 2.1 The space group $P2_1$ with the symmetry operators.

Braggs Law (real space)

Bragg demonstrated that a set of parallel planes with interplanar spacing d produces a diffracted beam where X-rays of wavelength λ impinge on the planes at an angle θ and are reflected at the same angle, only when θ meets the condition

$$n\lambda = 2d\sin\theta, \text{ where } n \text{ is an integer.}$$

Two rays R_1 and R_2 are reflected from two parallel planes, separated by the interplanar spacing d .

Lines AC are drawn from the point of reflection A of R_1 perpendicular the ray R_2 . If R_2 is reflected at B, then R_2 travels an extra distance of $2BC$ in relation to R_1 , which equates to $2d\sin\theta$ (**fig. 2.2**).

Braggs Law (reciprocal space)

An X-ray beam impinges on the crystal along the a^*b^* plane, passing through O along XO. O is the origin of the reciprocal lattice. The circle has radius $1/\lambda$, with centre C on XO. The reciprocal lattice point P lies on the circle. $OP = 1/d$ (**fig. 2.3**).

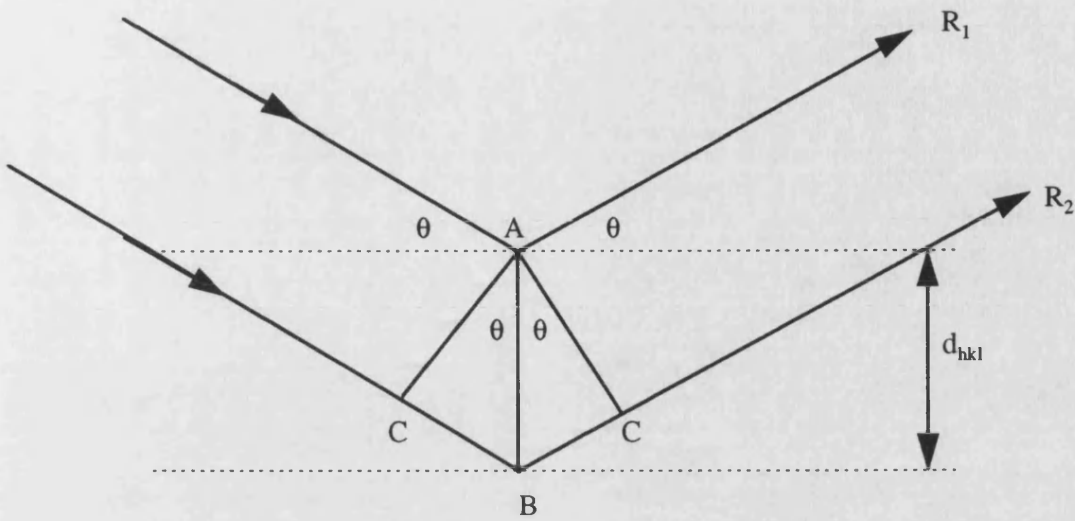


Fig. 2.2 Bragg's law. Two rays R_1 and R_2 , reflected at angle θ . Interplanar spacing = d_{hkl} . R_2 travels the extra distance $2BC$. $\sin\theta = BC/AB$, $BC = AB \sin\theta = d \sin\theta$.

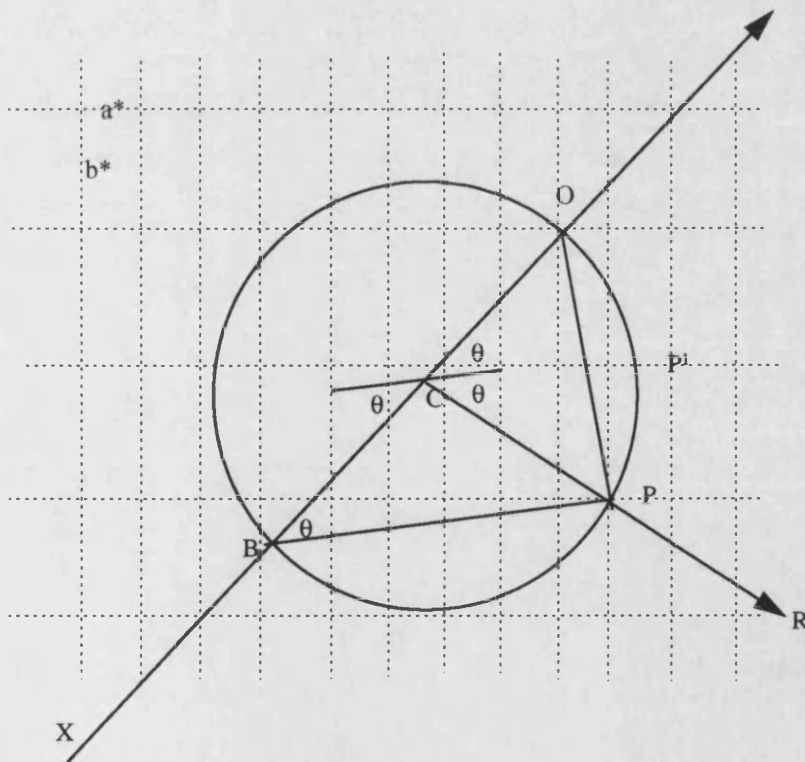


Fig. 2.3. Bragg's law in reciprocal space (a^* and b^*). P represents a reciprocal lattice point that intersects the circle, whereas P^1 does not. $\sin\theta = OP/BO = OP/2\lambda \Rightarrow n\lambda = 2d \sin\theta$.

The reflected X-ray occurs in the direction CP when P is in contact with the circle. From this, a sphere of reflection is established, whereby as the crystal is rotated in the X-ray beam, various reciprocal lattice points (P^1) come into contact with the sphere, each producing a beam in the direction of a line from the centre of the sphere of reflection through the reciprocal lattice point that is in contact of the sphere.

From this model of diffraction, the direction and number of reflections depend only on the unit cell dimensions. The intensity of the reflection hkl depends on the contents of the unit cell.

The Fourier series/transform

The Fourier transform describes the mathematical relationship that exists between the Fourier series description of an object (the electron density) and the Fourier series description of the diffraction pattern. The Fourier series of the reflection, the structure factor equation, contains a term for each atom in the unit cell.

A wave is a periodic function.

$$f(x) = F \cos 2\pi(hx + \alpha)$$

where $f(x)$ specifies the vertical height of a wave at any horizontal position x . F is the amplitude, h the frequency, and α the phase of the wave.

Any wave can be represented as a sum of simple wave components, that is wave can be represented as a Fourier series of n terms

$$f(x) = \sum_{h=0}^n F_h \cos 2\pi(hx + \alpha_h), \text{ which can be represented as}$$

$$f(x) = \sum_{h=0}^n F_h [\cos 2\pi(hx) + i \sin 2\pi(hx)] \text{ or}$$

$$f(x) = \sum_h F_h e^{2\pi i(hx)}$$

For a three dimensional wave, this is extended to

$$f(x, y, z) = \sum_{hkl} F_{hkl} e^{2\pi i(hx + ky + lz)}$$

The $f(xyz)$ function is related to the $f(hkl)$ by the Fourier transform, the operation of which is reversible (below). The Fourier transform relates real space to reciprocal space. $F(hkl)$ is the Fourier transform of $f(xyz)$, and similarly, $f(xyz)$ is the Fourier transform of $F(hkl)$.

$$F(hkl) = \int_{xyz} f(xyz) e^{2\pi i(hx + ky + lz)} dx dy dz$$

$$f(xyz) = \int_{hkl} F(hkl) e^{-2\pi i(hx + ky + lz)} dh dk dl$$

The structure factor, F_{hkl} , which describes a reflection, can also be represented as a Fourier series, in which each term gives the contribution of one atom to the reflection hkl . A single term in this Fourier series is represented as (a)

$$f_{hkl} = f_j e^{2\pi i(hx_j + ky_j + lz_j)} \quad (a)$$

$$F_{hkl} = \sum_{j=1}^n f_j e^{2\pi i(hx_j + ky_j + lz_j)} \quad (b)$$

$$F_{hkl} = \int_v \rho(xyz) e^{2\pi i(hx + ky + lz)} dv \quad (c)$$

The term f_j is the scattering factor of atom j . This function is slightly different for each element, as each element has a different number of electrons. x_j , y_j and z_j are the co-ordinates of atom j in the unit cell.

Each diffracted wave is a complicated wave, representing the sum of diffractive contributions from all atoms in the unit cell. The Fourier series for n terms is (b).

Alternatively, F_{hkl} can be represented as a sum of contributions from each volume element of electron density in the unit cell, where v is the unit cell volume. Here, F_{hkl} is the Fourier transform of the electron density $\rho(xyz)$ on the set of real lattice planes (hkl) (c).

The reversibility of the transform enables the electron density to be represented as a transform of the structure factors. Each term in the series, h, k , and l , are indices of the reflection hkl , and F_{hkl} is the structure factor, which is a complete description of a reflection. Each diffracted ray, or reflection, is a wave function that specifies frequency, amplitude and phase. The frequency is that of the X-ray

source, amplitude proportional to the square root of the intensity of the reflection, but the phase is unknown. This is the phase problem.

$$p(xyz) = \frac{1}{v} \sum_{hkl} F_{hkl} e^{-2\pi i (hx + ky + lz)}$$

$$|A| = |F| \cdot \cos \alpha, \quad |B| = |F| \cdot \sin \alpha$$

$$F = |F| \cdot (\cos \alpha + i \sin \alpha) \Rightarrow F = |F| \cdot e^{i\alpha}$$

$$p(xyz) = \frac{1}{v} \sum_{hkl} |F_{hkl}| e^{-2\pi i (hx + ky + lz - \alpha_{hkl})}$$

Representation of the structure factor as a complex vector enables a visualisation of the phase problem.

Isomorphous replacement

Each atom in the unit cell contributes to every reflection in the diffraction pattern, with the contribution of an atom being greatest when the indices of the reflection corresponds to the lattice planes that intersect the atom. Introduction of a small, electron dense number of atoms (a heavy atom) would cause changes in the intensity of each reflection of the diffraction pattern. This change in intensity enables an initial estimate of the phases (fig. 2.4 & 2.5).

To estimate the phases, the position of the introduced heavy atom needs to be determined. This is calculated using a Fourier series called the Patterson function $P(uvw)$, which is a sum of the interatomic vectors in Patterson space. The expression,

$$\Delta P(uvw) = \frac{1}{v} \sum_{hkl} \Delta F_{hkl}^2 \cdot e^{-2\pi i (hu + kv + lw)}$$

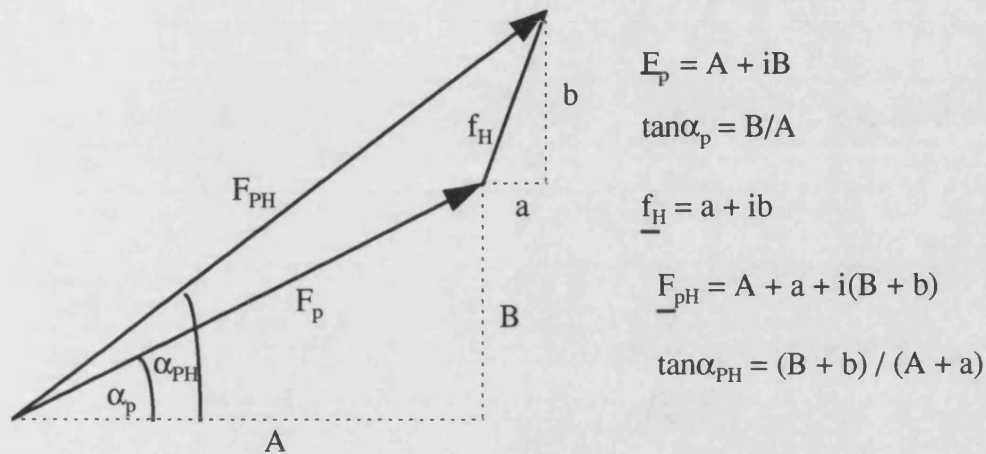
where $\Delta F^2 = (|F_{FH}| - |F_p|)^2$ in the difference Patterson function. The difference between the structure factor amplitudes \pm heavy atom reflects the contribution of the heavy atom only.

The search for the heavy atom in Patterson space is simplified by the unit cell symmetry. For $P2_1$, vectors connecting symmetry related atoms lie at $v = 1/2$, which is known as the Harker section. Two

solutions are obtained from the difference Patterson synthesis, which are mirror images of each other - this is where the phase ambiguity arises.

From knowing the phase of the heavy atom, the phases of all reflections can be estimated

$F_{PH} = F_P + f_H$. Fig. 2.4



The structure factor F_{hkl} has a real and imaginary component. The length of the vector (A) is the amplitude, whereas the phase represents the imaginary component. For centric reflections, the phase is 0° . F_p is the native structure factor, F_{PH} the derivative structure factor, and f_H the contribution of the heavy atom.

Usually, at least two heavy atom derivatives are required to overcome the phase ambiguity. The initial phases are only estimates, and therefore contain errors. Parameters such as phase probabilities, the figure of merit, closure errors are used to estimate the quality of the phases, and are used to weight the Fourier synthesis towards the more accurately determined phases.

Anomalous scattering : The ability of a heavy atom to absorb X-rays at a specific wavelength results in the break down of Friedel's law where the relationship $hkl \equiv -h -k -l$ does not hold. This inequality is a source of phase information, and can be sufficient to solve the phase problem.

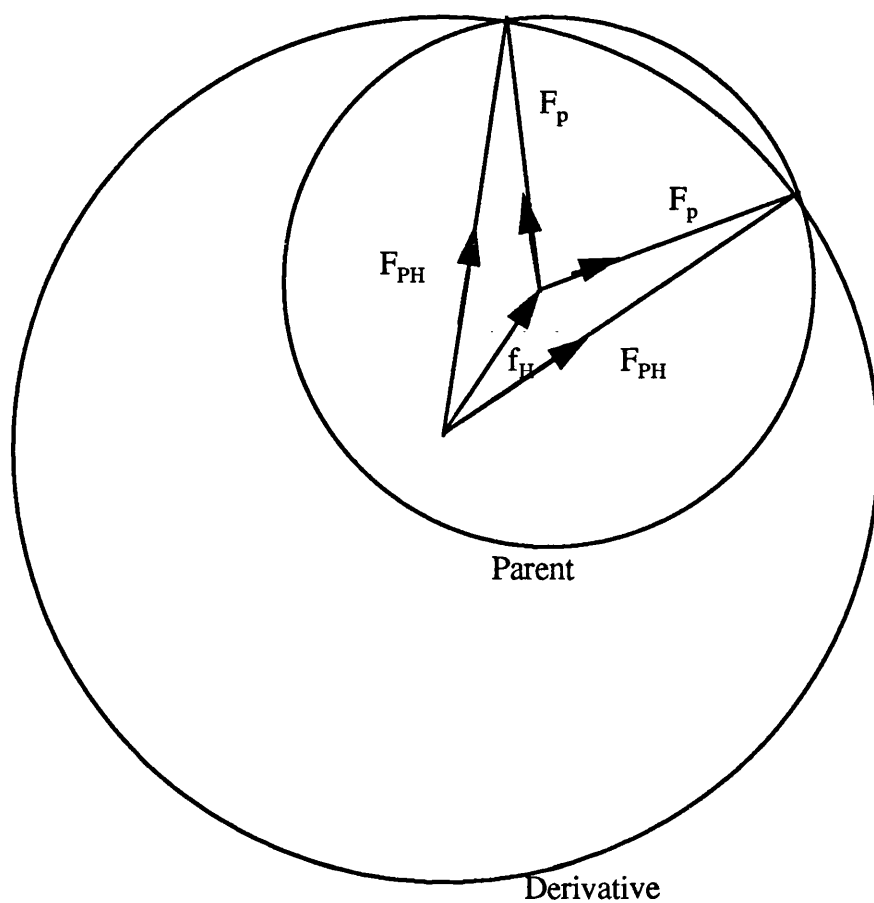


Fig. 2.5 Phase circle for single isomorphous replacement. The point of intersection represents the phase. For the SIR, there are two symmetrically related solutions. A second derivative,

Refinement

Crystallographic least squares refinement seeks to minimise the function

$$\Phi = \sum_{hkl} w_{hkl} (|F_o| - |F_c|)_{hkl}^2$$

where $|F_{obs}|$ is the observed structure factor amplitude and $|F_{calc}|$ is the calculated structure factor amplitude. Φ is the sum of the squares of differences between the observed and calculated amplitudes for all hkl reflections. The weight, w , is the reciprocal of the standard deviation of F_{obs} .

The positions of each atom j (x_j, y_j, z_j) determines each F_{calc} . The mobility of these atomic positions (the temperature factor, B_j) and the occupancy of each atom j (n_j) therefore affect the F_{calc} . The F_c term can be expanded to include these parameters

$$F_c = G \cdot \sum_j n_j f_j e^{2\pi i(hx_j + ky_j + lz_j)} \cdot e^{-B_j[(\sin\theta)/\lambda]^2}$$

G is an overall scale factor, f_j is the atomic scattering factor of atom j .

This function, Φ will exhibit many local minima, which corresponds to variations in the model, only one of which is correct. A least squares procedure finds the minimum that is nearest the starting point, thus the starting model must be near the global minimum. The greatest difference from the global minimum from which refinement will converge correctly is termed the radius of convergence ($d_{\text{min}}/4$). It follows that the greater the resolution of the model, the more accurate the starting model must be.

The radius of convergence can be increased by including constraints and restraints on the model during refinement. A constraint is a fixed value for a certain parameter (for example, setting all B -factors to 20\AA^2), while a restraint is a subsidiary condition imposed on the parameters, such as the bond length

$$\Phi = \sum_{\text{hkl}} w_{\text{hkl}} (|F_o| - |F_c|)_{\text{hkl}}^2 + \sum_i^{\text{bonds}} w_i (d_i^{\text{ideal}} - d_i^{\text{model}})^2 + \sum_j^{\text{angles}} w_j (\phi_j^{\text{ideal}} - \phi_j^{\text{model}})^2$$

where d_i is the bond length of i , and ϕ_j is the bond angle ϕ_j at location j

Abbreviations

1PGD	6-phosphogluconate dehydrogenase (Brookhaven code)
1SDG	Sorbitol dehydrogenase (Brookhaven code)
2GD1	Glyceraldehyde 3-phosphate dehydrogenase (thermophile) (Brookhaven code)
2GPD	Glyceraldehyde 3-phosphate dehydrogenase (Brookhaven code)
3GRS	Glutathione reductase (Brookhaven code)
4LDH	Lactate dehydrogenase (Brookhaven code)
6PGDH	6-phosphogluconate dehydrogenase
<i>A. calcoalceticus</i>	<i>Acinetobacter calcoalceticus</i>
ADP-ribose	Adenosine 5'-diphosphoribose
ATP-ribose	Adenosine Triphosphoribose (2'-Monophosphoadenosine 5' Diphosphoribose)
<i>B. megaterium</i>	<i>Bacillus megaterium</i>
<i>C. Symbiosum</i>	<i>Clostridium symbiosum</i>
DTT	dithiothreitol
EDTA	ethylene diaminetetra-acetic acid
G6PDH	glucose 6-phosphate dehydrogenase
GDH	glucose dehydrogenase
GLUTDH	Glutamate dehydrogenase
INFAC	Incomplete factorial
K _m	Michaelis constant
<i>L. Meristoidies</i>	<i>Leuconostoc mesenteroidess</i>
LADH	horse liver alcohol dehydrogenase
MIR	Multiple isomorphous replacement
NAD(P)	Nicotinamide adenine dinucleotide (phosphate)
NCS	Non-crystallographic symmetry
NMR	Nuclear magnetic resonance
pCMB	Parachloromercurybenzoate
PQQ	Pyroloquinoline quinone
rRNA	16s ribosomal RNA
SDH	sorbitol dehydrogenase
SDS	Soidum dodecyl sulhate
SIR	Single isomorphous replacement
SSADH	<i>Solfolobus solfataricus</i> alcohol dehydrogenase
<i>Tp. acidophilum</i>	<i>Thermoplasma acidophilum</i>
YADH	yeast alcohol dehydrogenase
Zn-ADH	long-chain zinc requiring alcohol dehydrogenase

References

- Abergel, C., Nesa, M.P. & Fontecilla-Camps, J.C. (1991). *J. Cryst. Growth*. **110**, 11-19.
- Abeysinghe, S.I.B., Baker, P.J., Rice, D.W., Rodgers, H.F., Stillman, T.J., Ko, Y. H., McFadden, B.A. & Nimmo, H.G. (1991). *J. Mol. Biol.* **220**, 13-16.
- Adams, M.J., Gover, S., Leaback, R., Philips, C. & Somers, D.O'N. (1991) *Acta Crystallogr. B* **47**, 817-820.
- Amaki, Y., Nakana, H., Yamane, T. (1994) *Appl. Microbiol. Biotech.* **40**, 664-668.
- Ammendola, S., Raia, C.A., Caruso, C., Camardella, L., D'Auria, S., De Rosa, M. & Rossi, M. (1992). *Biochemistry*. **31**, 12514-12523.
- Aronson, B.D. & Somerville, R.L. (1989) *J. Biol. Chem.* **264**, 5226-5232.
- Baker, P.J., Britton, K.L, Engel, P.C., Farrants, G.W., Lilley, K.S., Rice, D.W & Stillman, T.J. (1992). *PROTEiNS: Structure, Function and Genetics*. **12**, 75-86.
- Baker, P.J., Britton, K.L., Rice, D.W., Rob, A. & Stillman, T.J. (1992) *J. Mol. Biol.* **228**, 662-671.
- Barton, G.J. & Sternberg, M.J.E. (1990) *J. Mol. Biol.* **212**, 389-402.
- Barton, G.J. (1993) *Protein Eng.* **6**, 37-40.
- Blundell, T.L. & Johnson, L.N.(1976) In *Protein Crystallography*. Ed. Horecker, B., Kaplan, N.O., Marmur, J. & Scheraga, H.A. Academic Press Inc.
- Bohm, G. & Jaenicke, R. (1994) *Protein Engineering*. **7**, 213-220.
- Bork, P. & Grunwald, C. (1990). *Eur. J. Biochem.* **191**, 347-358.
- Borrás, T., Persson, B. & Jörnvall, H. (1989). *Biochemistry*. **28**, 6133-6139.
- Bricogne, G. (1976) *Acta Cryst.* **A32**, 832-846.
- Bright, J. R. (1991) *Ph. D. Thesis*. University of Bath.
- Bright, J. R., Mackness, R., Danson, M.J., Hough, D.W., Taylor, G.L., Towner, P. & Byrom, D. (1991). *J. Mol. Biol.* **222**, 143-144.
- Bright, J.R., Byrom, D., Danson, M.J., Hough, D.W. & Towner, P. (1993) *Eur. J. Biochem.* **211**, 549-554.

- Brünger, A.T. (1988) *J. Mol. Biol.* **203**, 803-816.
- Brünger, A.T. (1988) X-PLOR manual (version 2.1)
- Buehner, M., Ford, G.C., Moras, D., Olsen, K.W. & Rossmann, M.G. (1974) *J. Mol. Biol.* **82**, 563-585.
- Burley, S.K. & Petsko, G.A. (1985). *Science*. **275**, 23-28.
- Cammarano, P., Palm, P., Creti, R., Ceccarelli, E., Samanzantoni, A.M. & Tiboni, O. (1992). *J. Mol. Evol.* **34**, 396-405.
- Cammarano, P., Teichner, A. & Londei, P. (1986) *Syst. Appl. Microbiol.* **7**, 137-146.
- Campbell, D.P., Carper, W.R. & Thompson, R.E. (1982) *Arch. of Bioch. and Biophys.* **215**, 289-301.
- Carter, C.W., Baldwin, E.T. & Frick, L. (1988) *J. Cryst. Growth.* **90**, 60-73.
- Clermont, S., Corbier, C., Mely, Y., Gerard, D., Wonacott, A. & Branlant, G. (1993) *Biochemistry*. **32**, 10178-10184.
- Colonna-Cesari, F., Perahia, D., Karplus, M., Eklund, H., Brändén, C-I. & Tapia, O. *J. Biol. Chem.* **261**, 15273-15280.
- Danson, M.J. (1993) in *The Biochemistry of the Archaea (Archaeobacteria)*. Ed. Kates, M., Kushner, D.J. & Matheson, A.T. Elsevier, The Netherlands, **Chapter 1**, 1-21.
- Darland, G., Brock, T.D., Samsonoff, W. & Conti, S.F. (1970) *Science*. **170**, 1416-1418.
- De Rosa, M., Gambacorta, A., Minale, L., Thomson, R.H. & Worthington, R.D. (1977) *J. Chem. Soc. Perkin. Trans. 1*, 653-657.
- Ehrig, T., Hurley, T.D., Edenberg, H.J. & Borson, W.F. (1991) *Biochemistry*. **30**, 1062-1068
- Eklund, H. & Brandén, C-I. (1979). *J. Biol. Chem.* **254**, 3458-3461.
- Eklund, H., Horjales, E., Jörnvall, H., Brändén, C-I. & Jeffery, J. (1985). *Biochemistry*. **24**, 8005-8012.
- Eklund, H., Nordström, B., Zeppezauer, E., Söderlund, G., Ohlsson, I., Söderberg, B-O, Tapia, O., Åkeson, Å. & Brändén, C-I. (1976). *J. Mol. Biol.* **102**, 27-59.
- Eklund, H., Plapp, B.V., Samama, J-P. & Brändén, C-I. (1982). *J. Biol. Chem.* **257**, 14349-14358.
- Eklund, H., Samama, J-P. & Jones, T.A. (1984) *Biochemistry*. **23**, 5982-5996.

- Eklund, H., Samama, J-P., Wallén, L., Åkeson, Å., Jones, T.A. & Brändén, C-I. (1981) *J. Mol. Biol.* **146**, 561-587.
- Falchuk, K.H., Hilt, K.L. & Vallee, B.L. (1988). In *Methods of Enzymology*. Vol.158, Chapter 33, 422-434.
- Fersht, A.R. & Serrano, L. (1993). *Current Opinion in Structural Biology*. **3**, 75-83.
- Fitzgerald, P.M.D. (1988) *J. Appl. Cryst.* **21**, 273-281.
- Garret, R.A., Dalgaard, J.Z., Lykke-Andersen, J., Ostergaard, L. & Aagard, C. (1993). Symposium Abstracts: *Molecular Biology of the Archaea*. pg 31.
- Ghosh, D., Wawrzak, Z., Weeks, C.M., Duax, W.L., Erman, M. (1994) *Structure*. **2**, 629-640.
- Gogarten, J.P. & Taiz, L. (1992). *Photosynthesis research*. **33**, 137-146
- Gokhale, R.S., Agarwalla, S., Francis, V.S., Santi, D.V. & Balaram, P. (1994). *J. Mol. Biol.* **235**, 89-94.
- Grau, U.M. (1982). In *The Pyridine Nucleotide Coenzymes*. Ed. Everse, J., Anderson, B. & You, K-S.) Academic Press Inc., 135-187.
- Greer, J. (1974). *J. Mol. Biol.* **82**, 279-302.
- Hanukoglu, I. & Gutfinger, T. (1989). *Eur. J. Biochem.* **180**, 479-484.
- Henderson, P.J.F. (1992). In *Enzyme Assays, A Practical Approach*. Eds. Eisinger, R. & Danson, M.J. IRL Press. **Chapter 10**, 277-313.
- Hough, D.W. & Danson, M. J. (1989) *Lett. Appl. Microbiol.* **9**, 33-39.
- Hubbard, S.J., Gross, K-H & Argos, P. (1994). *Prot. Eng.* **7**, 613-626.
- Hurley, T.D., Borson, W.F., Stone, C.L. & Amzel, L.M. (1994) *J. Mol. Biol.* **239**, 415-429.
- Ikai, A. (1980). *J. Biochem.* **88**, 1895-1898.
- Iwabe, N., Kuma, K., Kishino, H., Hasegawa, M. & Miyata, T. (1991). *J. Mol. Evol.* **32**, 70-78.
- Jaenicke, R. (1991). *Eur. J. Biochem.* **202**, 715-728.
- Jancarik, J. & Kim, S-H. (1991). *J. Appl. Cryst.* **24**, 409-411.
- John, J., Crennell, S.J., Hough, D.W., Danson, M.J. & Taylor, G.L. (1994) *Structure*. **2**, 385-393.
- Jones, E.Y., Walker, N.P.C. & Stuart, D.I. (1991). *Acta Cryst.* **A47**, 753-770.

- Jones, T.A., Zou, J-Y, Cowan, S.W. & Kjeldgaard, M. (1991). *Acta Cryst.* **A47**, 110-119.
- Jörnvall, H., Eklund, H. & Brändén, C-I. (1978). *J. Biol. Chem.* **253**, 8414-8419.
- Jörnvall, H., Persson, B. & Jeffery, J. (1987) *Eur. J. Biochem.* **167**, 195-201.
- Jörnvall, H., von Bahr-Lindstrom, H. & Jeffery, J. (1984). *Eur. J. Biochem.* **140**, 17-23.
- Juhász, A., Csizmadia, V., Borbély, G., Udvardy, J. & Farkas, G.L. (1986) *FEBS.* **194**, 121-125.
- Kabsch, W. & Sander, C. (1983). *Biopolymers.* **22**, 2577-2637.
- Kabsch, W. (1988) *J. Appl. Cryst.* **21**, 916-924.
- Kandler, O. & König, H. (1993) in *The Biochemistry of the Archaea (Archaeobacteria)*. Ed. Kates, M., Kushner, D.J. & Matheson, A.T. Elsevier, The Netherlands, **Chapter 8**,
- Kates, M. (1993) in *The Biochemistry of the Archaea (Archaeobacteria)*. Ed. Kates, M., Kushner, D.J. & Matheson, A.T. Elsevier, The Netherlands, **Chapter 9**, 261-296.
- Kellis Jr, J.T., Nyberg, K., Šali, D. & Fersht, A.R. (1988). *Nature.* **333**, 784-786.
- Kleywegt, G.A. & Jones, T.A. (1994). In *Proceedings of the CCP4 study weekend*. Daresbury Laboratory, U.K. in press.
- Kleywegt, G.J. & Jones, T.A. (1994). *Acta Cryst.* **D50**, 178-185.
- Klinman, J.P. (1981). *CRC Crit. Rev. Biochem.* 39-78.
- Kraulis, P.J. (1991). *J. Appl. Crystallog.* **24**, 946-950.
- Laemmli, U.K. (1970). *Nature.* **227**, 680-685.
- Lake, J.A. (1991). *TIBS*, **16**, 46-50.
- Larsen, H. (1973) *Antonie van Leeuwenhoek J. Microbiol.* **39**, 383-396.
- Lee, C. & Levitt, M. (1991). *Nature.* **352**, 448-451.
- Luthy, R., Bowie, J.U. & Eisenberg, D. (1992). *Nature.* **356**, 83-85.
- Mackness, R. (1991) *Final Year Project*, University of Bath.
- Magonet, E., Hayen, P., Delforge, D., Delaive E. & Remacle, J. (1992) *Biochem. J.* **287**, 361-365.
- Mathews, B.W. (1968). *J. Mol. Biol.* **33**, 491-497.
- McPherson, A & Spencer, R. (1975). *Arch. Biochem. Biophys.* **169**, 650-661.
- McPherson, A. (1990). *Eur. J. Biochem.* **189**, 1-23.

- Menendez-Arias, L. & Argos, P. (1989). *J. Mol. Biol.* **206**, 397-405.
- Minor Jr, D.L. & Kim, P.S. (1994). *Nature*. **367**, 660-663.
- Morris, A.L., MacArthur, M.W. & Thornton, J.N. (1992). *Proteins*. **12**, 345-364.
- Nazava, J. (1992) AMoRe. In Molecular replacement. (Eds. Dodson, E.J., Gover, S. & Wolf, W.) SERC Daresbury Laboratory, Warrington, UK. 87-90.
- Ohlsson, I., Nordström, B. & Brändén, C-I. (1974) *J. Mol. Biol.* **89**, 339-354.
- Olsen, G.J. & Woese, C.R. (1993). *FASEB J.* **7**, 113-123.
- Otwinowski, Z. (1993) In (Sawyer et al., 1993) To appear, ISSN 0144-5677.
- Pal G.P., Jany K-D & Saenger, W. (1987) FEBS. 123-124
- Ramaswamy, S., Eklund, H. & Plapp, B.V. (1994) *Biochemistry*. **33**, 5230-5237.
- Rayment, I. (1983). *Acta Cryst.* **A39**, 102-116.
- Rayment, I. (1985) *In Methods in Enzymology*. **Vol.114**, Chapter 10.
- Rhodes, G. (1993) *In Crystallography Made Crystal Clear*. Academic Press Inc.
- Richardson, J.S. & Richardson, D.C. (1988). *Science*. **240**, 1648-1652.
- Rivera, M.C. & Lake, J.A. (1992) *Science*. **257**, 74-76.
- Rossmann, M.G. (1990). *Acta Cryst.* **A46**, 73-82.
- Rossmann, M.G., Moras, D. & Olsen, K.W. (1974) *Nature*. **250**, 194-199.
- Russell, R.J.M. & John, J. (1994) *In Archae and Extremophiles newsletter*.
- Schlunegger, M.P., Grütter, M.G., Streiff, M.B., Olsthroon, A.J.J. & Duine, J.A. (1993) *J. Mol. Biol.* **233**, 784-786.
- Scrutton, N.S., Berry, A. & Perham, R.N. (1990) *Nature*. **343**, 38-43.
- Searcy, D.G. (1986) *Syst. Appl. Microbiol.* **7**, 198-201.
- Smith, L. (1989). *Ph. D. Thesis*, University of Bath.
- Smith, L.D., Budgen, N., Bungard, S.J., Danson, M.J., Hough, D.W. (1989) *Biochem J.* **261**, 973-977.
- Stetter, K.O. & Zillig, W. (1985) in *'The Bacteria, vol. 8: The Archaeobacteria'* ed. C.R. Woese and R.S. Wolfe. Academic Press Inc., New York.

- Stillman, T.J., Baker, P.J., Britton, K.L., Rice, D.W. & Rodgers, H.F. (1992) *J. Mol. Biol.* **224**, 1181-1184.
- Story, R. & Steitz, T.A. (1992). *Nature*. **355**, 374-376.
- Stura, E.A. & Wilson, I.A. (1991) *In Crystallization of Nucleic Acids and Proteins, A Practical Approach*. Ed. Ducruix, A. & Giegé, R. **Chapter 5**, 99-126.
- Sun, H-W. & Plapp, B.V. (1992). *J. Mol. Evol.* **34**, 522-535.
- Sund, H. & Theorell, H. (1963). *The Enzymes* (2nd ed.) **7**, 25-83.
- Sutcliffe, M.J., Haneef, I., Carney, D. & Blundell, T.L. (1987). *Protein Eng.* **1**, 377-384.
- Tête-Favier, F., Rondeau, J-M, Podjarny, A. & Moras, D. (1993). *Acta Cryst.* **D49**, 246-256.
- Thaller, C., Eichele, G., Weaver, L.H., Wilson, E., Karlsson, R. & Jansonius, J.N. (1985). *In Methods in Enzymology*. **Chapter 9**, 132-135.
- Thaller, C., Weaver, L.H., Eichele, G., Wilson, E., Karlsson, R. & Jansonius, J.N. (1981) *J. Mol. Biol.* (1981) **147**, 465-469.
- Theorell, H. & Tatemoto, K. (1971). *Arch. Biochem. And Biophys.* **142**, 69-82.
- Theorell, H. Chance, B. (1951) *Acta Chem. Scand.* **5**, 1127-1144.
- Tonrud, D.E., Ten Eyck, L. & Mathews, B.W. (1987) *Acta Cryst.* **A43**, 489-501.
- Tsai, C.S., Senior, D.J. & Al-Kassim, L.S. (1989). *Comp. Biochem. Physiol.* **94B**, 655-659.
- Vallee, B.L. & Auld, D.S. (1990) *Biochemistry*. **29**, 5647-5659.
- Varghese, J.N., Laver, W.G. & Colman, P.M. (1983) *Nature*. **303**, 35-40.
- Verlinde, C.L.M.J. & Hol, W.G.J. (1994) *Structure*. **2**, 577-587.
- Wang, B-C. (1985). *In Methods of Enzymology*. Academic press. Inc. **Vol. 115**, Chapter 7, 90-113.
- Wharton, C.W. & Eisenthal, R. (1981) *In Molecular Enzymology*. **Chapter 7**, 169-203.
- Wierenga, R.K., De Maeyer, M.C.H. & Hol, W.G.J. (1985). *Biochemistry*. **24**, 1346-1357.
- Woese, C.R., Kandler, O. & Wheelis, M.L. (1990). *Proc. Natl. Acad. Sci. USA*. **87**, 4576-4579.
- Woese, C.R., Olsen, G.J. (1986) *Syst. Appl. Microbiol.* **7**, 161-177.
- Zhang, K.Y.J. (1993). *Acta Cryst.* **D49**, 213-222.

References

Zhang, Z., Djebli, A., Shoham, M., Frolow, F., Peretz, M. & Burstein, Y. (1993). *J. Mol. Biol.* **230**, 353-355.

Zillig, W., Klenk, H-P, Palm, P., Leffers, H., Pühler, G., Gropp, F. & Garrett, R.A. (1989) *Endocytobiosis & Cell Res.* **6**, 1-25.