Single crystal x-ray diffraction studies on small, medium and large molecules

Susanna Butterworth

The work described in this thesis was carried out at Durham and Hamburg between October 1992 and August 1996, under the supervision of J.A.K.Howard. Unless otherwise stated it is my own work and has not been submitted previously for a degree at this or another university.

SINGLE CRYSTAL X-RAY DIFFRACTION STUDIES ON SMALL, MEDIUM & LARGE MOLECULES

Susanna Butterworth

Abstract

Chapter 1. Production of crystals for diffraction analysis would be assisted by the devising of a set of rules which, given molecular formula, could predict crystal formation conditions. By studying trends in structural properties of a group of closely related simple molecules, deductions could be drawn which could then be applied more generally. Chalcone derivatives with minor substituent differences were recrystallised, X-ray diffraction data collected and the structures solved and refined. Additionally, NMR and UV studies were performed, investigating an observed dimerisation reaction.

Chapter 2. Discovery of peptide hormones and neurotransmitters has stimulated the study of structure-activity relationships, although the structure of these molecules is often poorly defined. Proctolin, a linear pentapeptide, is a neurotransmitter in insects. Crystallisation was attempted, with the aim of deducing the active conformation structure, thereby assisting in design of small molecule analogues for use as non-cholinergic pesticides. No diffraction was observed from the crystals produced.

Chapter 3. Glucosamine 6-phosphate synthase is an N-terminal nucleophile amidotransferase catalysing the first step in the hexosamine pathway, from which all amino-sugar containing macromolecules are derived. Structure determination of each of two subdomains was attempted. In one case, pseudo-symmetry appeared to obstruct structure solution. The symmetry has subsequently been understood and the structure obtained. Crystals of the second domain are rotationally disordered.

Chapters 4 and 5. Recent advances in macromolecular crystallographic techniques have facilitated the collection of an increasing number of high quality, atomic resolution data sets. Methods for refinement, previously limited to small molecule structures, have increasing relevance for proteins. Atomic resolution refinements using these evolving protocols have been performed on two small proteins, rubredoxin from *Desulfovibrio vulgaris* and the protein G immunoglobulin-binding domain. Appropriate treatment of the solvent structure in a protein crystal and the benefit to be gained by using sharpened density maps during refinement were investigated.

Acknowledgements

I would like to thank everyone I have known in Durham and Hamburg for help, encouragement and laughs along the way. In particular I thank my supervisor, Judith Howard, for putting up with me and assisting me through all my changes of direction. I am very grateful to Keith Wilson for making it possible for me to come to Hamburg in the beginning and for his consistent support through my time there. This thesis is dedicated to Jon, who has shared my life through all the ups and downs of the last four years.

Single Crystal X-ray Diffraction Studies

on

Small, Medium and Large Molecules

SUSANNA BUTTERWORTH

DEPARTMENT OF CHEMISTRY, UNIVERSITY OF DURHAM

submitted in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September, 1996

Table of Contents

General Introduction

I began writing an account of my work during the last four years by drawing up a plan of the topics the thesis should cover. The unusually wide range of subject matter seemed to demand an introductory explanation of how the PhD had evolved, setting out my ambitions and goals at each stage, a summary of my achievements and experiences, and an explanation of why the direction of study changed and how my aims were modified over the course of time.

The project, entitled "diffraction studies and NMR investigation of medium-sized molecules" was embarked upon in the framework of a CASE award between Shell Research agrochemical division in Sittingbourne, Kent and the University of Durham Chemical Crystallography Department, commencing in October 1992. The goal of the Shell research program was the development of a new generation of non-cholinergic pesticides, the initial object being to obtain a structural model of the active conformation of an oligopeptide, proctolin, a substance with neuroregulatory activity in many species of insects and some other invertebrates as an input for modelling studies to suggest possible small molecular analogues. NMR studies on the material in progress at Shell had proved rather inconclusive, although I had little access to this information at that stage, so single crystal X-ray diffraction studies were required. Shell funded a supply of the material from Sigma Chemicals to the Chemical Crystallography Department in Durham. I attempted to obtain crystals. Realising that this was a different scale of problem from the recrystallisations of small molecules I had previously performed, I sought the advice of experts: Marek Brzozowski at York University and Steve Wood at Birkbeck College, London.

Crystallisation attempts were undertaken without positive result until April 1993, when I accepted the generous offer of Steve Wood to spend some time under his supervision at Birkbeck learning about the practical techniques entailed in the crystallisation of peptides and larger molecules. In the following months at Birkbeck, some limited success was achieved, with the production of crystals from which no diffraction could be observed.

In September 1993 the research project at Shell came to an end, with the take-over by a rival company and subsequent closure of the department in Sittingbourne. Shell Research, while legally obliged to continue funding the CASE award, had no further scientific interest in the project.

I returned to Durham and decided to continue work on a project which had originally been an Oxford Chemistry Part II project with David Watkin at the Chemical Crystallography Laboratory. This was a study of the properties of a set of closely related small organic compounds, chalcones, with the aim of deriving a relation between variation in formula and crystallisation behaviour and the nature of the crystal lattice. This work had been initially successful but much remained to be done and I had regretted leaving it unfinished at the end of a year. At Durham, further recrystallisation, data collection, structure solution, NMR and UV studies were performed on the chalcones, along with the re-refinement of the previously obtained structures, in the light of my increasing crystallographic knowledge. A body of results accumulated but I was still at loss to know how to interpret them and derive relations between chemical formulae and lattice properties. Inspiration was required and it did not come. As the project stood, it did not constitute a PhD thesis.

In the spring of 1994 I attended the HERCULES School in Grenoble on the use of large facilities in solid state studies, for which I had applied while still working on proctolin. Here I learnt more about applications of synchrotron and neutron sources, especially in the field of biological macromolecular structures. The course provided new motivation and perspective and I realised I was getting stuck in a rut.

Shortly afterwards, I visited the EMBL synchrotron outstation at DESY, Hamburg. Discussions with Keith Wilson led to the suggestion that I might be able to spend a period working at the synchrotron, where there were projects available, given that I still had 18 months funding. This proposal was agreed by Judith Howard in Durham and by the EPSRC and in June 1994 I arrived in Hamburg.

At Hamburg I began working with Alex Teplyakov on the determination of the structure of two subdomains of an enzyme glucosamine 6-phosphate synthase. Crystals had already been obtained for both domains and initial data collection was underway. The project provided me with an introduction to the practical techniques of protein crystallography including crystallisation, synchrotron data collection, searching for heavy atom derivatives and molecular replacement. Unfortunately, progress soon became hindered by the appearance of crystal symmetry problems. The pseudo-symmetry of the crystals of one of the domains appeared to obstruct structure solution. My work on this domain was continued by Misha Usupov at Exeter University, who was eventually successful in gaining an understanding of the extremely subtle and complex nature of the crystal symmetry and so solving the structure. The crystals of the other domain were found to possess a kind of rotational disordered lattice. My attempts to somehow obtain crystals with better symmetry were fruitless. No significant progress in tackling this problem has since been made.

In January 1995 it was decided that, with only nine months remaining before the end of the third year of the PhD and no guarantee of imminent results, it would be advisable to do some work of a more analytical nature. For the next fifteen months I worked in collaboration with Victor Lamzin on the refinement of atomic resolution protein structures and the development of tools for the analysis and refinement of protein crystal structures. Thus, I increased my knowledge of some more theoretical areas of crystallography and gained an introduction to FORTRAN and C-Shell programming. And finally obtained some more tangible results.

Looking at the overall thesis plan, it is possible to see a theme running through the series of projects covered. My initial experience in crystallisation concerned the essentially trivial solution and refinement of some small molecular structures with the far from trivial aim of developing a systematic method for the prediction of crystallisation conditions. I then went on to work on a larger molecule, a pentapeptide, thereby gaining first-hand experience of the problems and frustrations of the search for crystallisation conditions. In my next project, the molecules were considerably bigger, several hundred amino acids in length. In this case, crystallisation conditions had been found, the problem lay with the next step, structure solution, a different class of problem for a protein from that presented by running a default direct methods job on a twenty atom structure. My final objects of study were slightly smaller, with around sixty residues each. In the development of techniques for the treatment of atomic resolution protein data sets I was concerned with the progress of protein crystallography into the traditional territory of small molecular crystallographers

- almost (but not quite) completing the circle. Thus, the entire thesis could be seen as progress towards mapping out a continuum of structures, from molecules comprising a few atoms to those made up of hundreds of kilodaltons.

Abbreviations used in this thesis

people & places

DESY	Deutsches Elektronen-SYnchrotron
EMBL	European Molecular Biology Laboratory
EPSRC	Engineering and Physical Sciences Research Council
PDB	Brookhaven Protein Data Bank (Bernstein *et al.*, 1977)
CSD	Cambridge Structural Database (Allen *et al.*, 1991)

crystallographic terminology, especially concerning proteins

Å	1 Ångström = 01 nm
θ	diffraction angle
λ	wavelength of radiation (commonly X-ray) (Å)
d	resolution of diffracted intensity, Braggs law, $d = \lambda / 2\sin\theta$ (Å),
s	resolution expressed as $2\sin\theta / \lambda$ (Å^{-1})
F	structure factor
F_o / F_c	structure factor amplitude; observed / calculated
E	normalised (sharpened) structure factor
H	half-sharpened structure factor, $F^{½}E^{½}$ (defined for the purposes of this thesis)
α_c	calculated phase of a reflection
R	reliability index, $\Sigma \lvert \lvert F_o \rvert - \lvert F_c \rvert \rvert / \Sigma \lvert F_o \rvert$
R_{free}	R factor evaluated from a subset of data independent of refinement (Brünger, 1993)
B	isotropic thermal parameter, $8\pi^2 \langle u \rangle^2$ (Å^2), where $\langle u \rangle^2$, root mean square displacement of atom
ρ	electron density, commonly expressed in e Å^{-3}
V_m	crystal volume per unit molecular weight (Matthews, 1968) (Å^3 D^{-1})
k_{sol}	contrast parameter for modelling diffuse solvent, $\Sigma\rho_{sol}(r) / \Sigma\rho_{protein}(r)$ (Tronrud, 1996)
B_{sol}	average B factor for diffuse solvent (Å^2) (Tronrud, 1996)
rms(d)	root-mean square (distance/deviation, commonly in Å)
H-bond	hydrogen bond, A_{donor}-H···$A_{acceptor}$, where donor and acceptor atoms commonly O,N,S possibly C
NCS	non-crystallographic symmetry

naturally occurring amino-acids

Ala (A)	alanine	Leu (L)	leucine
Arg (R)	arginine	Lys (K)	lysine
Asn (N)	asparagine	Met (M)	methionine
Asp (D)	aspartate	Pro (P)	proline
Cys (C)	cysteine	Phe (F)	phenylalanine
Gly (G)	glycine	Ser (S)	serine
Glu (E)	glutamate	Thr (T)	threonine
Gln (Q)	glutamine	Trp (W)	tryptophan
His (H)	histidine	Tyr (Y)	tyrosine
Ile (I)	isoleucine	Val (V)	valine

substances concerned with crystallisation

PEG	polyethylene glycol (precipitant)	DMSO	dimethyl sulphoxide
TFA	trifluoroacetic acid	DMF	dimethyl formamide
DTT	dithiothreitol (antioxidant)	MPD	methyl-pentanediol *(additive)*
HEPES	2-[4-(2-hydroxyethyl)-1-piperazino]-ethanesulphonic acid (buffer)		
TRIS	α,α,α-tris-(hydroxymethyl)-methylamine (buffer)		
EMTS	mercury thiosalicylate		

substances concerned with biochemical processes

ATP / ADP / AMP	adenosine triphosphate / diphosphate / monophosphate
GTP	guanosine triphosphate
UDP	uridine diphosphate
$NADP^+$	nicotinamide adenine dinucleotide phosphate oxidised
NADPH	nicotinamide adenine dinucleotide phosphate reduced
Ig / IgG	immunoglobulin / immunoglobulin G
Pi / PPi	inorganic orthophosphate / pyrophosphate

other techniques

HPLC	high performance liquid chromatography
NMR	nuclear magnetic resonance

Chapter 1:

A systematic study of the crystal structures of phenyl acetophenones

Background

References to phenyl acetophenones, or chalcones, have occurred in the literature of various fields. An investigation into the polymorphism displayed by these compounds was reported by Weygand (1929). A study of relationships between structure and solid state reactivity of conjugated organic molecules, including chalcones, was carried out at the Weizmann Institute of Science. The effect of crystal packing on photopolymerisation reactions was investigated (Rabinovich & Schmidt, 1970) and stereospecific syntheses, enabled by the crystallisation of non-chiral molecules into chiral spacegroups, were studied (Rabinovich & Shakked, 1974). The pharmacological activity of several chalcones has been investigated, for example, the use of kukulkanins, which can be extracted from the bark of *Mimosa tennuefolia*, by Mexican Indians to heal burns (Dominguez *et al.,* 1989). A recently reported study of chalcones was concerned with the relationship between the crystal structure, which is non-centrosymmetric and the non-linear optical properties displayed by the material (Zhiengdong *et al.*, 1992).

Introduction

Single X-ray crystal diffraction is the choice method for molecular structure determination for molecules of all sizes. The first hurdle to be overcome in such an analysis is the production of crystals of suitable quality for diffraction studies. Despite almost a century of experience, little is known about how and why a substance crystallises in a particular way from a specific set of conditions. The formulation of a set of rules which, given the molecular formula, could be applied to predict conditions for the crystal formation of a substance would be extremely useful (Van der Sluis, 1989). One direction of approaching this goal is to study the crystal properties of a group of simple small molecules with very closely related structures. Deductions drawn from systematic trends traced within the group could then be generalised and applied across a wider field.

The opportunity for a systematic study of the relationships between the molecular structure of a compound and its crystalline properties was offered by the synthesis of a series of twenty-five organic compounds with only minor differences in substituents on a common backbone: chalcone derivatives with varying numbers of oxy-methyl and methylene-dioxy substituents on the two aromatic rings, Table 1. A number of these chalcones were recrystallised under varying conditions, most commonly by the evaporation of a toluene or an ethanol solution, and X-ray diffraction data were collected on a total of seventeen different types of crystal. Details of these experiments are summarised in the Appendix at the end of the chapter.

A survey of crystallisation conditions and the determination of twelve structures were carried out before the start of this PhD (Bahar, 1992). During this PhD, eleven of the structures were re-refined and the details of these refinements are presented, along with the four structures determined since, at Durham, in the Appendix. The chalcone XII crystals were extensively disordered, preventing determination of the spacegroup. No final conclusions were reached about this structure. Similar problems were experienced with chalcone XV, following data collection on these crystals at Durham. Crystals of the remaining ten chalcone molecules could not be grown under the conditions successful for the first fifteen.

Table 1. Methoxy-chalcones synthesised

	12	13	14	15	16	2	3	4	5	6	I.d	structures determined
2-	*	-	*	-	-	-	-	-	-	-	I	1 [a]
MeO	-	-	*	-	-	-	-	*	-	-	II	1 [a]
	-	-	-	-	-	-	*	*	-	-	III	0
	-	-	*	-	-	-	*	-	-	-	IV	0
	*	-	-	-	-	-	-	*	-	-	V	0
3-	-	-	-	-	-	-	*	*	*	-	VI	1 [d]
MeO	*	-	*	-	-	-	-	*	-	-	VII	0
	-	-	-	-	-	*	-	*	-	*	VIII	0
	*	-	*	-	*	-	-	-	-	-	IX	0
	-	*	*	-	-	-	-	*	-	-	X	1 [a]
	-	-	*	-	-	-	*	*	-	-	XI	1 [a]
4-	*	-	*	-	*	-	-	*	-	-	XII	1 [b]
MeO	*	-	*	-	-	-	*	*	-	-	XIII	0
	-	-	*	-	-	-	*	*	*	-	XIV	1 [a]
	-	*	*	*	-	*	-	-	-	-	XV	1 [e]
	-	-	*	-	-	*	-	*	-	*	XVI	1 [d]
5-	*	-	*	-	*	-	*	*	-	-	XVII	1 [a]
MeO	-	*	*	-	-	-	*	*	*	-	XVIII	0
	*	-	*	-	*	-	*	-	*	-	XIX	1 [a]
	-	*	*	*	-	-	*	*	-	-	XX	0
MeO	-	-	-	-	-	-	↔		-	-	XXI	1 [a]
/	*	-	*	-	-	-	↔		-	-	XXII	2 [f]
O	-	↔		-	-	-	↔		-	-	XXIII	0
- CH$_2$-	-	↔		-	-	-	*	*	-	-	XXIV	2 [c]
O	*	-	*	-	*	-	↔		-	-	XXV	1 [a]

[a] structure determined at Oxford (Bahar, 92), re-refined at Durham

[b] structure determined at Oxford (Bahar, 92), extensively disordered

[c] structures of 2 polymorphs determined at Oxford (Bahar, 92), re-refined at Durham,

[d] structure determined at Durham

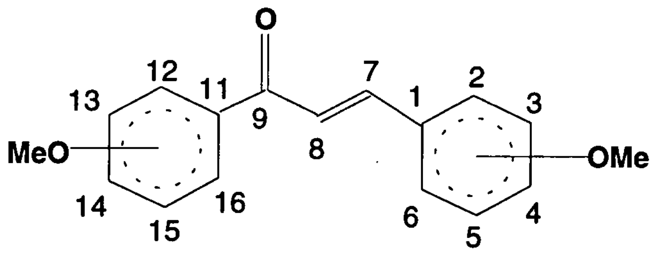
[e] data collected at Durham, extensively disordered

[f] monomer and dimer structures determined at Durham

* methoxy group substituted at this position

↔ methylene-dioxy group substituted at these 2 positions

Numbering scheme for substitution positions on the chalcone molecule



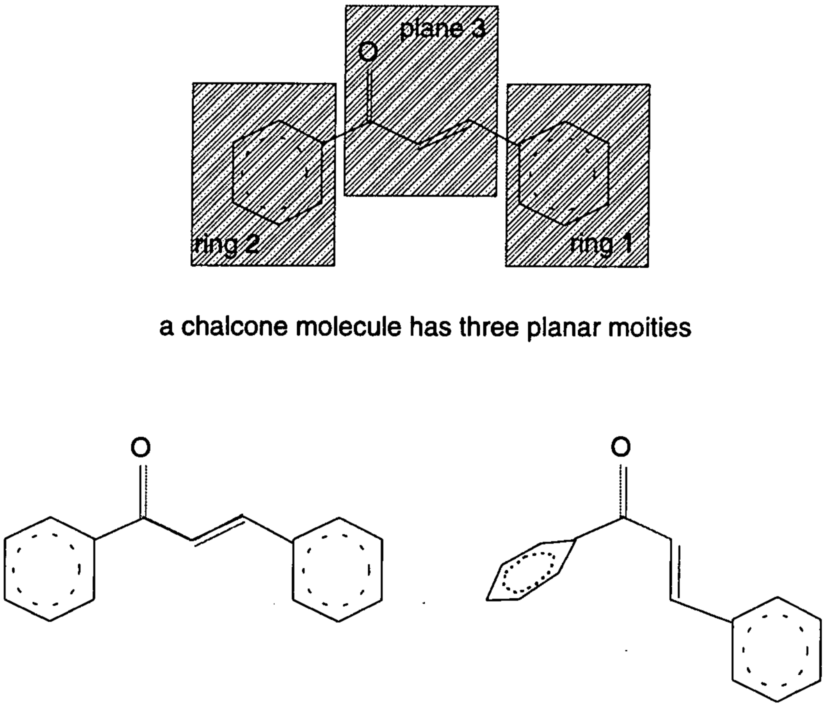
The first part of this chapter is a survey of the molecular conformations and packing motifs adopted in thirteen chalcone structures and the relationships between crystal properties and molecular formula. The final part comprises an account of the discovery and investigation of the dimerisation reaction of chalcone XXII.

Molecular conformation

The molecule possesses three approximately planar parts. The dihedral angle between ring 2 and plane 3 falls into one of two ranges: 0-35° and 85-90°, corresponding to the two conformational types, "cisoid" and "transoid", Figure 1. The dihedral angle ring 1 / plane 3 lies in the range 0-16°. These parts of the molecule are almost coplanar in all instances.

The molecular conformation is determined by the balance of two sets of competing intramolecular forces: stabilisation due to π-π overlap across the three planar regions of the molecule and destabilisation due to steric repulsion between C(12) and C(16) substituents and O(1) and H(8) and between C(2) and C(6) substituents and H(8) and H(7).

Figure 1, Molecular conformation



a chalcone molecule has three planar moities

cisoid conformation

transoid conformation

The stabilisation due to conjugation is maximised when the entire molecule is planar. However steric repulsion is also greatest in this conformation. Because of the position of the ketone oxygen, the repulsion experienced by ring 2 when coplanar with the central part is greater than that experienced by ring 1 and consequently the stabilising and repulsive forces between ring 2 and plane 3 are more closely balanced. When hydrogens are present at C(12) and C(16), ring 2 lies coplanar, but when methoxy groups are substituted, ring 2 becomes perpendicular to the rest of the molecule. Ring 1 remains coplanar with plane 3 even when methoxy groups are substituted at C(2) and C(6), the lowering in energy provided by the conjugation of ring 1 with the unsaturated ketone being the dominating stabilising force.

Packing characteristics - a survey of fourteen chalcone crystal structures

Molecular conformation affects the nature of the packing within the crystal. Molecules in transoid conformation have a crossed shape, which limits their packing possibilities, preventing stacking in sheets or the interlocking herringbone pattern commonly seen in the structures of flat aromatic species. The flat cisoid conformation lacks these restrictions. Consequently the appearance of more unusual packing motifs and distinctive features, such as disorder and the incorporation of solvent, might be expected in the transoid molecule lattices. The crystal structures have been divided into small groups for description and comparison.

Chalcones I, XVI & XXI *(Table 9a)*

All three crystallise in space group *Pbca* and the unit cell dimensions for the three structures are roughly similar, viz., 7, 14, 28Å.

In the chalcone I structure, sheets of molecules run perpendicular to the *b* axis. The ring 1 phenyl groups of adjacent molecules form an interlocking herringbone pattern. Ring 2 forms parallel stacks with approximately 40% overlap between adjacent rings, with an interplanar separation of 3.66 Å. The division of the sheets into two regions, one of coplanar packed rings the other with adjacent rings aligned in a herringbone fashion, can be seen in the view down the *a* axis, Figure 2a.

Chalcone XXI also forms rippled sheets with the plane perpendicular to the *b* axis. There are interactions between pairs of molecules, with the ketone of one molecule lying in proximity to the dioxy-methylene ring of its partner, Figure 2b. The sheet is divided into bands running parallel to the *a* axis, aromatic sectors occupied by phenyl groups and

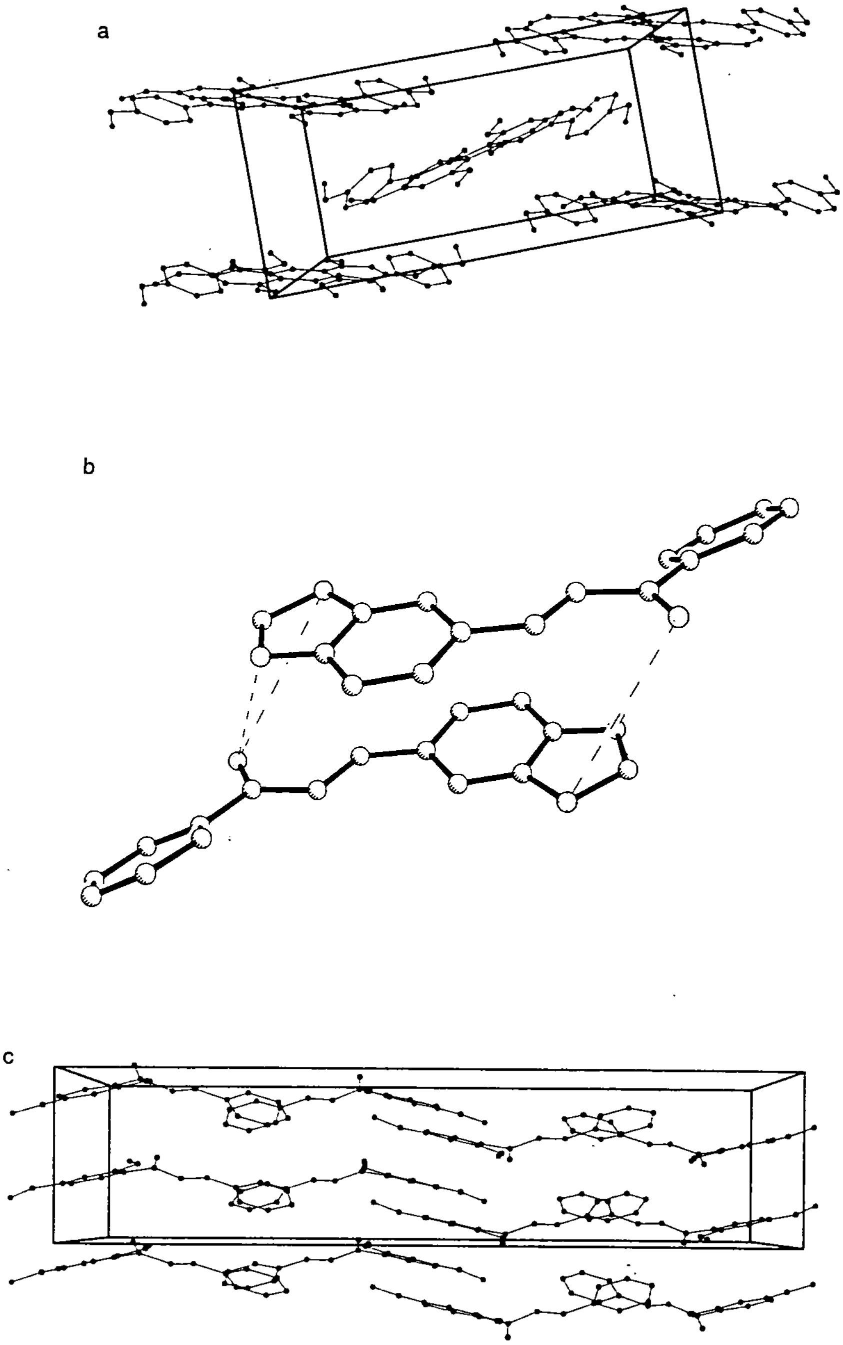


Figure 2. Structures I, XVI and XXI

a/ The packing diagram of the I structure, viewed down the *a* axis, shows the division of puckered sheets of molecules into π-π bonding and herringbone interlocking regions.

b/ An illustration of the pairing of molecules in the XXI structure with close contacts formed between the methylene dioxy oxygens of one molecule with the ketonic oxygen of the other.

c/ Packing diagram of the XVI structure, viewed down the *b* axis. There are no sheets in this structure.

polar sectors in which the oxygen atoms lie. The molecules are aligned in pairs within the sheet. According to Kitaigorodsky (1961), the packing of eight molecules in a unit cell in *Pbca* often implies that they are closely bound in pairs, as seen here.

Chalcone XVI crystallises in the same space group as I and XXI, but the latter two structures have more features in common. In the I and XXI structures, molecules are arrayed in rippled sheets perpendicular to the short axis, Figure 2c. In the chalcone XVI structure the plane of the molecules is roughly parallel to the short axis of the unit cell, but the pattern of rippled sheets is absent and it is less easy to divide the lattice up into regions of different chemical environment. The aromatic rings lie parallel but do not overlap.

Chalcones II, X & XI *(Table 9b)*

Chalcones II and XI crystallise in non-centrosymmetric space groups, $P2_12_12_1$ and $P2_1$ respectively. The molecules are of similar shape, essentially planar with a twist about the long axis of the molecule, resulting in a chiral conformation. There are examples (Rabinovich & Shakked, 1974), in which the crystallisation of achiral chalcones in non-centrosymmetric space groups in this fashion has allowed solid state asymmetric substitution on the olefinic bond.

The modes of packing of chalcones II and XI are very similar, stacking parallel to the short *a*-axis, Figure 3a. There is no direct overlap between aromatic rings on adjacent molecules within the stacks. The planes of rings of molecules in neighbouring columns are tilted relative to one another. Chalcone X adopts an entirely different kind of lattice from the other two, Figure 3b. The molecules form planar sheets perpendicular to [1 1 0].

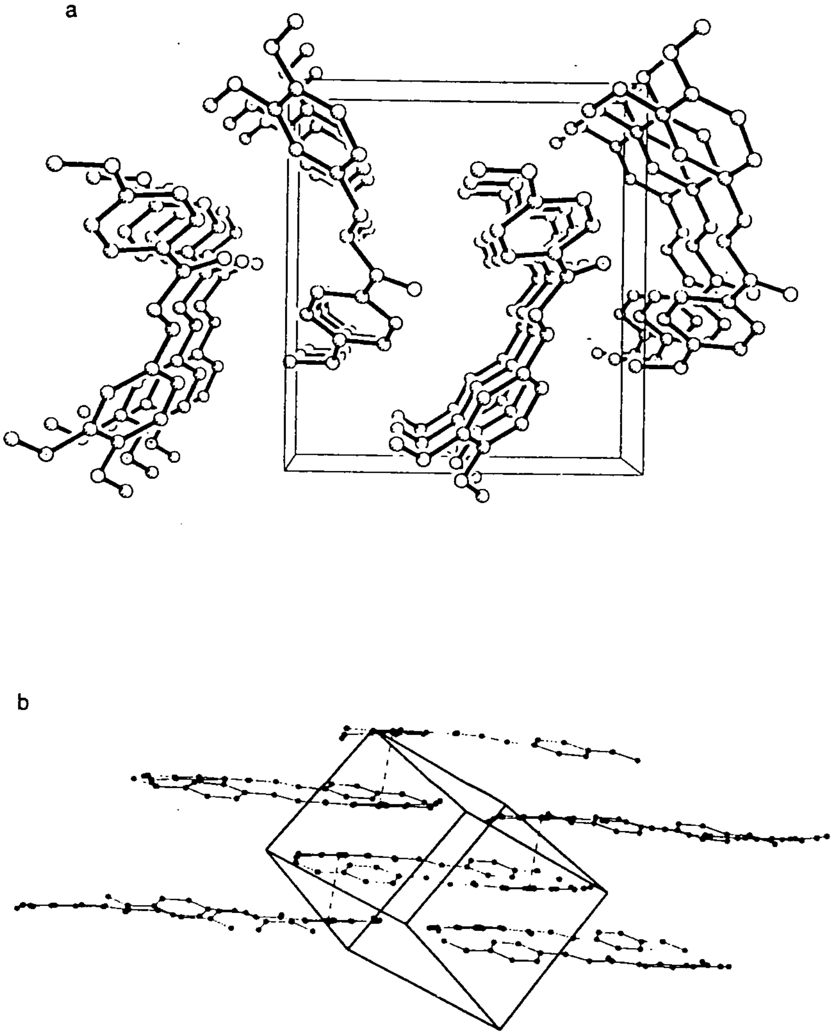


Figure 3. Structures II, X and XI

a/ Packing diagram for the XI structure, viewed down the a axis.

b/ Packing diagram for the X structure, viewed parallel to the molecular plane, illustrating the associating of the molecules in pairs within sheets and the plane-plane overlap of rings of molecùles in neighbouring sheets.

Within sheets the molecules are associated in pairs. Plane-plane overlap of one of the rings occurs between molecules in adjacent sheets with an interplanar distance of 3.5 Å.

An extra methoxy group added to the meta position on ring 1 of chalcone II produces chalcone XI, which has a crystal structure with a similar packing arrangement to that of chalcone II. Conversely, if the methoxy group is added to the meta position on ring 2, chalcone X is produced, which adopts a markedly different packing motif. The reason for this difference is not obvious.

Chalcones XXIV and XXV *(Table 9c)*

Comparison of the two polymorphs of chalcone XXIV demonstrates the effect of π–π interactions in the crystal lattice of a small organic molecule. Both XXIV(A) and XXIV(B) structures are in space group $P\ \bar{1}$. However, there are two molecules in the asymmetric unit of XXIV(A) giving Z=4, while there is a single molecule in the asymmetric unit of XXIV(B).

The two independent molecules in the XXIV(A) lattice adopt notably different conformations, as can be seen from the differences in dihedral angles, Table 2. The two molecules have different functions in the structure. The planar molecule 1 is aligned in stacks running along [1 0 1], the molecular plane lying perpendicular to the stacking direction. There is an inversion centre between each pair of molecules in the stack and plane-plane overlap between the rings at each end, with an interplanar separation of 3.5 Å. Molecule 2 lies with its molecular plane perpendicular to [1 2 0]. There is no interplanar overlap between aromatic rings. From the packing diagram viewed down [1 0 1], Figure 4a, it can be seen that molecule 2 acts as a space-filler between the molecule 1 stacks.

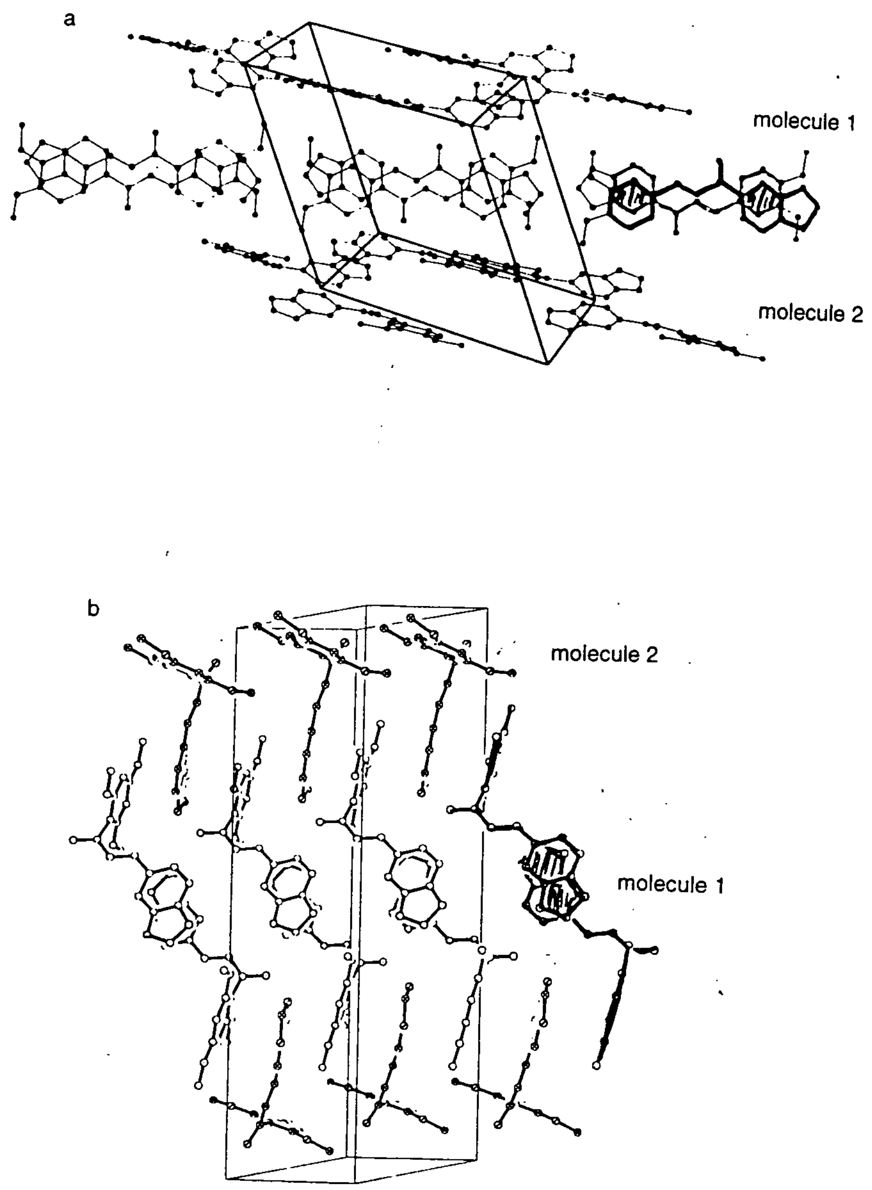


Figure 4. Structures XXIVA, XXIVB and XXV

a/ Packing diagram for the XXIVA structure, viewed down [1 0 1], illustrating the different environments ot the two molecules.

b/ The packing diagram of the XXV structure looking in the [1 -1 0] direction shows that the two molecules in the unit cell have different functions in this example as well.

Table 2. Dihedral angles for chalcones XXIV and XXV

Molecule	dihedral angle (°) plane 1/plane 3*	dihedral angle (°) plane 2/plane 3*	dihedral angle (°) plane 1/ plane 2*
XXIV(A), mol 1	8	0	8
XXIV(A), mol 2	6	26	32
XXIV(B)	5	6	7
XXV, mol 1	13	86	76
XXV, mol 2	16	88	78

* using notation of Figure 1

The driving force behind crystallisation of the compound in the XXIV(A) form is the lowering in lattice energy afforded by π–π interactions along the molecule 1 stacks, with a second molecule incorporated to fill spaces between the stacks, its conformation distorted to fit the shape of the cavity.

The second polymorph, XXIV(B), has a lattice characterised by planar layers of molecules, with no overlap between rings of molecules in adjacent layers. The stacks of overlapping aromatic molecules in the XXIV(A) crystal structure and their absence in the XXIV(B) lattice can be directly related to the marked difference in colour between the two types of crystal: XIV(A) being bright yellow, while XXIV(B) is colourless. This contrast can be ascribed to the stabilising π-π interactions in the XXIV(A) lattice, which cause a shift to longer wavelength of the UV- chromophore.

The chalcone XXV molecule has substituents at both *ortho* positions on ring 2 and adopts the transoid molecular conformation, as discussed above, while all the molecules so far discussed have cisoid conformation. Despite the difference in molecular conformation, the chalcone XXV crystal structure has several features in common with XXIV(A).

Chalcone XXV crystallises in $P\,\bar{1}$, with two molecules in the asymmetric unit. Although the two molecules adopt similar conformations, their environments are different, Figure 4b. Molecule 1 forms paired layers, which lie perpendicular to the *c* axis. Ring 1 of each molecule projects into the space between the layers, overlapping with ring 1 extended from the parallel layer, to form π–π interacting stacks, which run through the lattice down [1,-1,0] with an interplanar separation of 3.58 Å between the overlapping rings. Molecule 2 acts as a space-filler in the lattice.

It is possible that chalcone XXV could crystallise in a second polymorph with Z=2 and without π–π interactions, but no evidence of this form exists.

Chalcones XVII & XIX *(Table 9d)*

Chalcones XVII & XIX both adopt a transoid conformation, since both ortho-positions on ring 2 are substituted. Members of the set that have the planar, cisoid conformation are likely to adopt modes of packing commonly seen among aromatic small organic molecules. Alignment of the molecule in a transoid conformation, in which ring 2 lies perpendicular to the plane of the rest of the molecule, leads to the formation of more uncharacteristic types of lattice.

The chalcone XVII structure is unique among those of the methoxy chalcones so far investigated in being the only one of the set to incorporate solvent molecules within the lattice. The structure is less dense than the others and the space group, *P 2/a,* is unique in the group and less common generally. The crystallisation in a lattice with this space group can be related to the presence of the water molecules around the two-fold rotation axes. The stability of this structure can be attributed to the formation of H-bonds with the carbonyl oxygen O(1) acting as an acceptor and hydrogen atoms of the water molecules as donors. The lattice is divided into hydrophilic and hydrophobic regions. H-bonds traverse the polar zone, linking pairs of ketonic oxygen atoms via a water molecule, Figure 5a.

Ring 1 stacks running parallel to the *a* axis make up apolar regions in the lattice with π–π interactions between neighbouring rings in the stack. Two types of overlap alternate along the stack, one between molecules related by an inversion centre, with an interplanar separation of 3.46 Å, the other between molecules related by a rotation axis, with an interplanar separation of 3.49 Å. Ring 2, which has a dihedral angle of 78° with ring 1, forms a border between polar and apolar regions in the lattice, Figure 5b.

The packing motif in the chalcone XIX structure bears some resemblance to that of chalcone XVII. The cross shaped molecules are aligned in pairs in a similar fashion to that seen in the chalcone XVII lattice. However, in the XIX lattice, the stacks of ring 1 only form π-π interactions between discrete pairs of overlapping rings, with an interplanar separation of 3.37 Å. There is no polar solvent layer between the aromatic bands.

Chalcone VI ***(Table 9e)***

The chalcone VI molecule is essentially planar and is packed in herringbone fashion, a feature common to several of the other cisoid chalcone structures investigated and which is commonly seen in the crystal structures of flat, aromatic molecules.

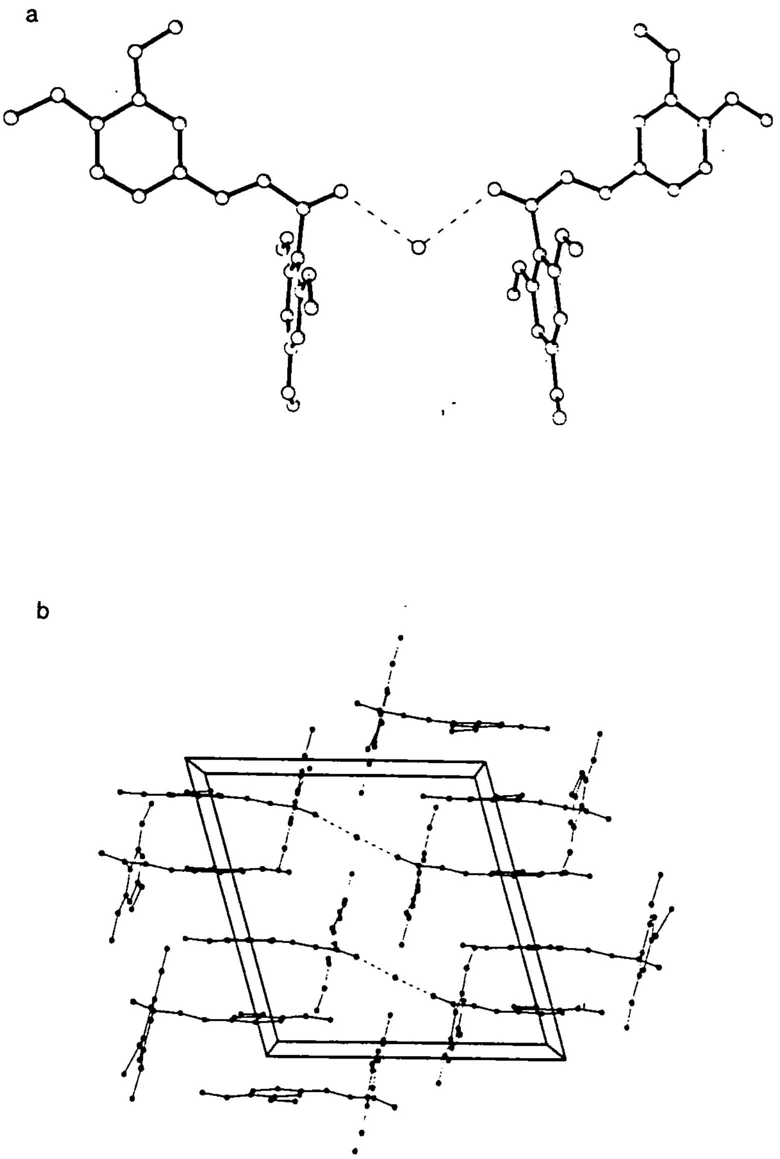


Figure 5. Structures XXVII and XIX

a/ 2-fold related molecules in the XVII structure, their ketonic oxygens linked via a water molecule on the axis.

b/ The packing in the XVII structure, illustrated by the view down the *b* axis. The lattice is divided into apolar regions, containing stacks of overlapping aromatic rings, and polar, H-bonding regions.

Chalcone XIV *(Table 9e)*

The chalcone XIV molecule adopts a twisted conformation in the crystal lattice. The molecules are associated in pairs around an inversion centre. Stacks of molecules run parallel to the *a* axis. There is no direct overlap between the aromatic rings of adjacent molecules in a stack. It is not clear whether crystal field forces in the lattice of chalcone XIV cause the twisted molecular conformation, or if this is a low-energy solution conformation which crystallises to give this particular mode of packing.

Conclusions to packing survey

The crystal structures possess a range of notable features. These include:

- Crystallisation of a compound in two polymorphs: one form containing two molecules in the asymmetric unit, the other, only one molecule.
- Stacks of overlapping aromatic rings in several of the structures, suggesting the occurrence of extensive π-π bonding.
- Inclusion of waters of crystallisation in one example, allowing the formation of H-bonded chains through the lattice.
- Crystallisation of two structures in asymmetric space groups, although the molecules are inherently achiral, thus producing two enanteomorphic forms of crystal, which could not be differentiated in this case.
- Disorder in two of the structures studied.

Relationships were traced between crystallographic and molecular properties, notably that between the position of the methoxy substituents and the molecular conformation adopted in the crystal structure. However, the reason behind the specific arrangement of a particular molecule in a crystal lattice is not obvious. A prediction of the molecular conformation of members of this set, for which structures were not obtained, can be made based on the positions of methoxy groups on the molecule. However, other features of the crystal lattice cannot be predicted reliably.

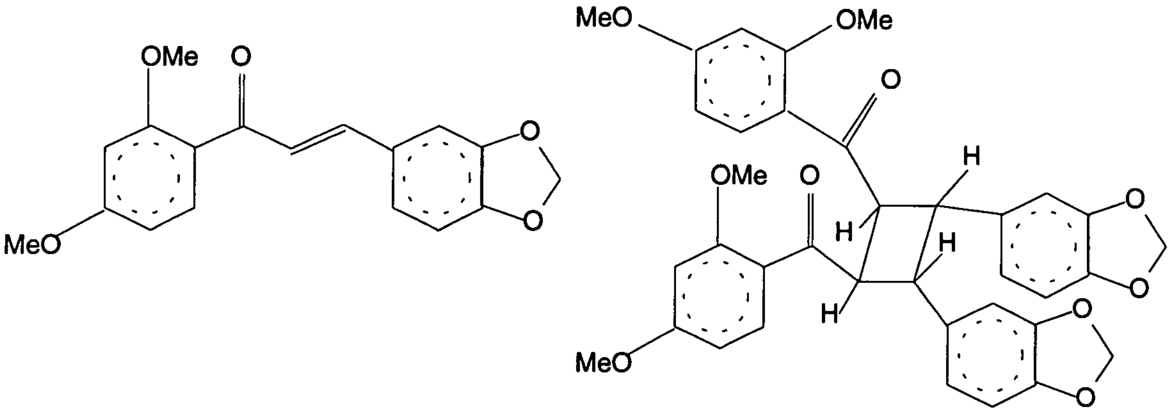
Dimerisation of chalcone XXII

Summary

The investigation of chalcone XXII initially yielded the crystal structure of a dimer. Subsequently, crystals of the monomer were also obtained. In the monomer structure the molecules are aligned in an orientation which would allow photocatalysed dimerisation with the stereospecific production of the stereoisomer of the dimer crystal. ^{13}C NMR spectra provide evidence that only one stereoisomer has been formed. The path by which the formation of the dimer occurred remains uncertain. However the existence of the monomer crystal structure described gives a strong indication that it is a solid-state reaction.

The original sample, with formula reported as in Figure 6 (1), was recrystallised by vapour diffusion, using toluene as the solvent and petroleum ether as the precipitant. The crystals grew in the form of small, monoclinic, pale yellow blocks. X-ray diffraction studies on these crystals showed that they were only weakly diffracting. Data were collected on a Siemens rotating anode. The structure was solved and refined, proving to be that of the dimer Figure 6 (2).

Figure 6. Chalcone XXII monomer and dimer structures



(1) monomer

(2) dimer

^{13}C NMR spectroscopy was employed to probe the composition of the original sample, in the solid state and in acetone solution. This showed that the dimerisation reaction was stereospecific, resulting in formation of a single isomer of the dimer.

Crystals grew in an acetonic solution of the original sample, at the bottom of an NMR tube. Unlike those previously obtained from toluene solution, the new crystals were large, bright yellow needles. X-ray diffraction data were collected on the second form of crystal using a Rigaku AF6S diffractometer. The structure was solved and refined, proving to be that of the monomer.

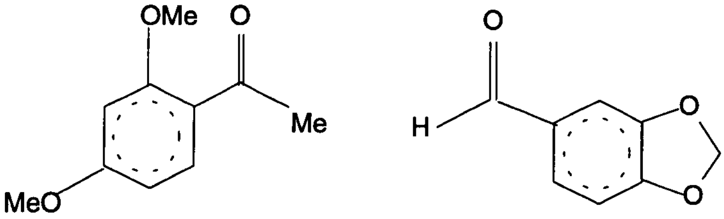
Synthesis of fresh monomer was performed as described below. ^{13}C NMR spectroscopy demonstrated that the sole product of the synthesis was the monomer.

Crystals of chalcone XXII monomer were photographed on a Weissenberg camera, irradiated with light of λ=350 nm for 90 minutes and then photographed again, in an attempt to detect a solid state dimerisation reaction. An acetone solution of the freshly synthesised chalcone XXII monomer was irradiated with light λ=380 nm for 40 minutes. The ^{13}C NMR spectrum was then recorded to investigate whether dimerisation, or other reactions, are catalysed in the solution by ultra-violet light. No dimerisation or other chemical or structural change could be detected by these experiments, in either solution or crystalline state.

Synthesis of chalcone XXII

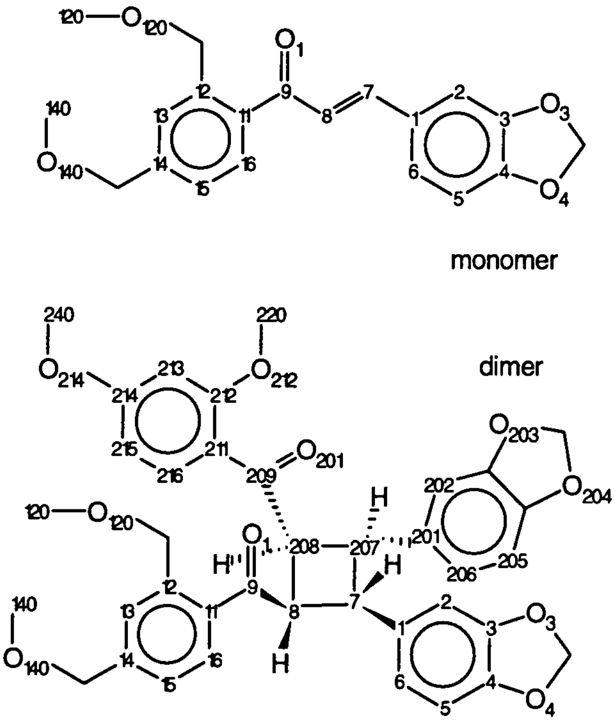
0.1 ml 2,4 dimethoxy acetophenone, Figure 7 (1) and 0.1 ml piperonal Figure 7 (2) were charged to a conical flask supported on a stirrer/hot plate. 200 ml of ethanol was added and the flask blanketed with nitrogen. After a few minutes stirring the reactants dissolved. Once the solution was clear, sodium hydroxide pellets (0.5 g) were added. The reaction mixture was then stirred under nitrogen until T.L.C. analysis indicated that the reaction has stopped. This took about six hours. After filtration, the crude product was recrystallised twice from ethanol. The pure product was then dried.

Figure 7. Reactants in chalcone XXII synthesis



(1) (2)

Figure 8. Numbering scheme for XXII monomer and dimer



Crystal Structure Determination

Crystals of the monomeric compound used in X-ray diffraction experiments grew from acetone solution in the form of large, bright yellow needles. Data were collected on a crystal of dimensions 0.9x0.5x0.4 mm. Crystals of the dimer, small, pale yellow, monoclinic blocks, were grown using vapour diffusion with toluene as the solvent and petroleum ether as the precipitant.

For the monomer crystal, cell dimensions and intensity data were measured on a Rigaku AFC6S diffractometer. Data collection on the Rigaku proved the dimer crystal to be only weakly diffracting and larger crystals were not available, giving good reason to refine cell parameters and collect intensity data using a Siemens rotating anode. Other details of the refinements are summarised in the Appendix and Table 9f.

Discussion of crystal structures

monomer crystal

The unit cell has an extremely short, 3.91 Å, *a* axis. Stacks of molecules, lying in the *bc* plane, run along the short axis, Figure 9c. There is complete coplanar overlap between adjacent molecules in the stack, Figure 9d, thus double bonds of neighbouring molecules are in alignment for photo catalysed dimerisation, Figure 10, (Woodward & Hoffmann, 1970), the distance between the planes formed by C(1), C(7), C(8), C(9) & O(1) of neighbouring molecules being 3.74 Å. Dimerisation from this configuration would result in stereospecific production of the isomer found in the dimer crystal structure. Photodimerisation often occurs in the solid state where double bonds on adjacent molecules are aligned in this manner, Figure 11.

There is a number of short intermolecular C-H...O distances, in the range 3.1-3.8 Å, between molecules lying in a plane, Table 3. The presence of these links suggests that the lattice is held together by two principal types of intermolecular interaction, π-π bonding along stacks running parallel to (1, 0, 0) and C-H...O bonding between neighbouring stacks.

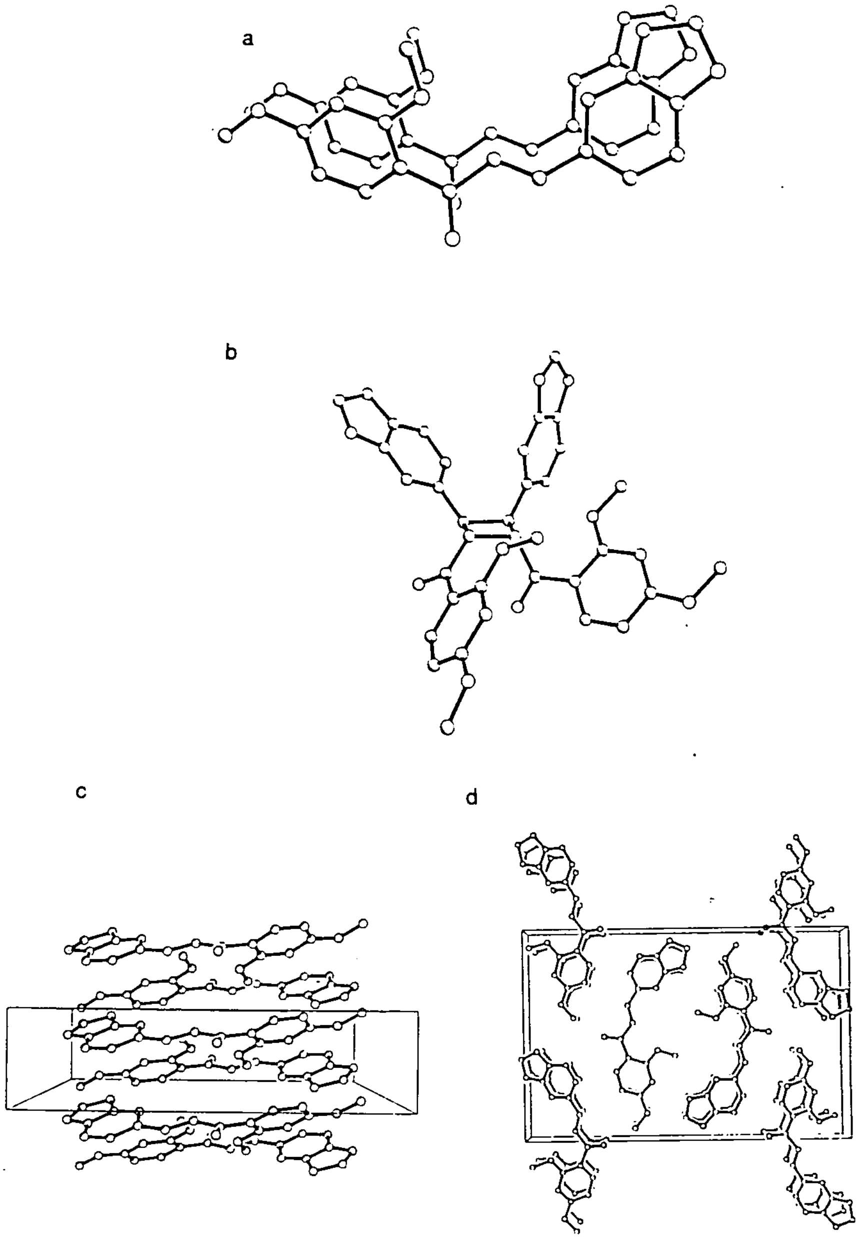


Figure 9. The chalcone XXII monomer and dimer structures

a/ The alignment of neighbouring molecules in the monomer lattice permits the occurrence of a solid-state dimerisation.

b/ The dimer molecule does not possess mirror symmetry in the solid state.

c/ The packing diagram for the monomer structure viewed down the *c* axis illustrates the close proximity of neighbouring molcules in stacks along the *a* axis.

d/ The view of the XXII structure down the *a* axis depicts the arrangement of molecules in a sheet, and the short C-H···O contacts within the sheet.

Table 3. Short C-H···O intermolecular distances in the monomer crystal structure

O(1) - C(140)	3.11
O(3) - C(120)	3.57
O(4) - C(140)	3.76
O(12) - C(120)	4.02
O(14) - C(43)	3.78

Table 4. Interplanar angles in the monomer crystal structure

angles between planes	
1 & 3	8.4°
2 & 3	37.9°
1 & 2	43.9°

Figure 10.
Intermolecular distances between C=C bonds in the XXII monomer structure

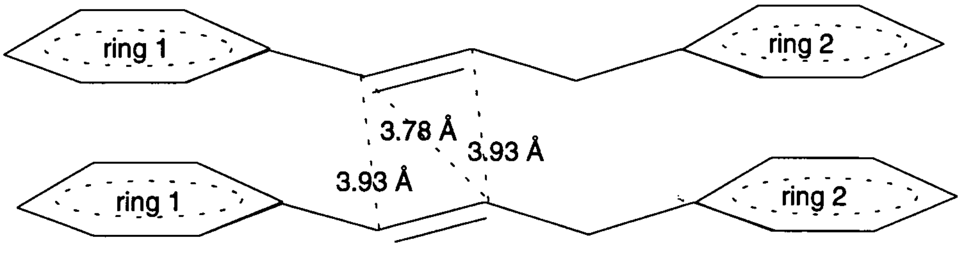
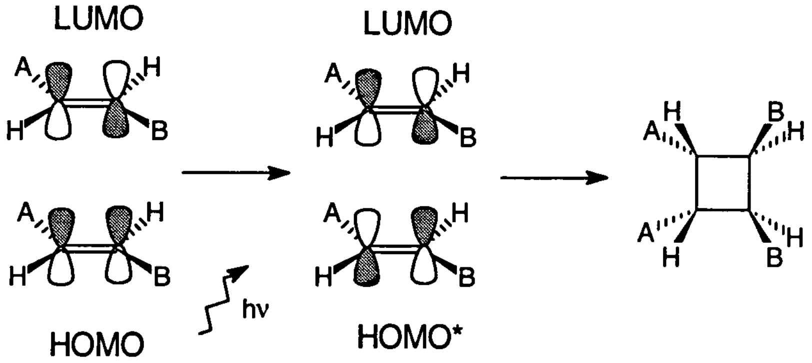


Figure 11. Dimerisation catalysed by UV radiation



dimer crystal

The conformation and alignment of the molecules in the lattice bear no relation to those in the monomer structure, as illustrated in Figure 9a & b and by the values of dihedral angles, Table 5. This does not prove the reaction does not occur in the solid state. All that is certain is that the reaction occurred before the dimer was recrystallised from the acetone solution, in solution, or previously, in a solid phase.

Planes in the XXII dimer

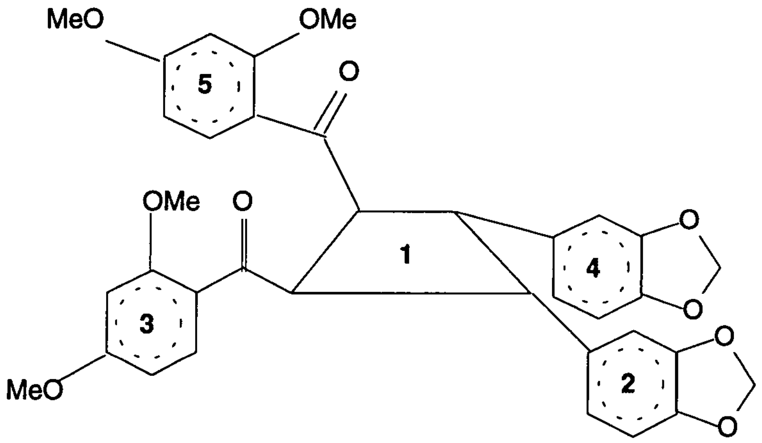


Table 5. Dihedral angles, XXII dimer

angle between planes		angle between planes	
1&2	93.4°	2&3	93.4°
1&3	54.3°	2&4	49.1°
1&4	44.3°	2&5	122.3°
1&5	52.7°	3&4	65.3°
4&5	84.3°	3&5	29.7°

^{13}C NMR studies on chalcone XXII

Solution spectra were recorded on a Varian VXR 400-S spectrometer in D_6 acetone solutions. Solid-state spectra were recorded on a Varian VXR 300 Solid-State spectrometer. Peak assignment for both monomer and dimer molecules were carried out using three solution spectra:

1/ freshly synthesised monomer

2/ original sample, composed of a mixture of monomer and dimer

3/ original sample, following removal from solution of a proportion of the monomer by crystallisation.

Peaks in the pure monomer spectrum were assigned with the aid of a predicted spectrum obtained from a ^{13}C chemical shift calculation by SPECINFO (Daresbury).

The assignment of dimer peaks in the mixed spectra could then be carried out, using three approaches:

1/ subtraction of peaks assigned to the monomer using the pure monomer spectrum

2/ comparison of the two dimer spectra - in the second spectrum, the dimer peaks become relatively more intense than the monomer peaks

3/ use of a spectrum predicted for the dimer molecule by SPECINFO.

The results of this peak assignment are listed in Table 6

Table 6. Assignment of peaks in ^{13}C solution spectra

monomer spectrum (ppm)	assignment C(No.)	prediction (ppm)	**dimer** spectrum (ppm)	assignment C(No.)	prediction (ppm)
189.82	9	190.0	198.01	9	191.6
165.13, 161.36	4, 2*	164.4, 162.8	165.19, 161.28	4, 2*	164.4, 160.4
150.36, 149.34	13, 14*	147.8, 147.5	148.13, 146.31	14, 13*	147.5, 147.3
141.68	7	143.6			
133.09	6	132.3	136.14	11	131.9
130.83	11	129.2	133.13	6	128.3
126.37	16	122.8	121.97, 121.33	1, 16*	126.9, 121.9
125.50	8	119.8			
123.05	1	122.2			
109.25, 107.22, 106.53	5, 15, 12*	110.5, 108.3, 107.8	109.33, 108.23, 106.46	12, 5, 15*	111.2, 110.5, 109.1
102.57	43	101.1	101.51	43	101.1
99.11	3	98.1	98.42	3	98.1
56.18, 55.91	20, 40*	56.0, 55.4	55.84, 55.74	20, 40*	56.0, 55.4
			53.42, 45.19	8, 7*	52.0, 44.9

* these assignments could not be made more specific and are interchangeable.

Two solid-state spectra were obtained: one from the sample first received containing a mixture of monomer and dimer and one on a freshly synthesised pure monomer sample. Seventeen peaks were visible in the spectrum of the pure monomer which could be assigned by comparison to the solution spectrum of the pure monomer.

The spectrum of the original sample is more complex. It contains monomer peaks, as seen in the spectrum of pure monomer, as well as dimer peaks, which are split due to the removal of the mirror-plane through the cyclo-butane ring in the solid-state and, in addition, a third carbonyl peak. All the other peaks in the spectrum can be assigned to either monomer or dimer. Assignments of solid-state spectra are set out in Tables 7 & 8.

Table 7. Assignments for Solid-State Spectrum of Pure Monomer

Solution spectrum (ppm)	Assignment C(no)	Solid-state spectrum (ppm)
189.8	9	188.9(N)
165.1, 161.4	4, 2*	163.4(N), 157.4(N)
150.4, 149.3	13, 14*	149.6(N), 147.89(N)
141.7	7	138.0(P)
133.1	6	133.0(P)
130.8	11	129.3(N)
126.4	16	128.6(P)
125.5	8	124.5(P)
123.1	1	123.1(N)
109.3, 107.2, 106.5	5, 15, 12*	108.6(P), 103.2(P)**
102.6	43	101.3(P)
99.1	3	99.3(P)
56.2, 55.9	20, 40*	57.0(P), 55.1(P)

(P) - Protonated carbon

(N) - Non-Protonated carbon

* these assignments could not be made more specific and are interchangeable.

** only two out of three peaks visible

Table 8. Assignment of Solid-State Spectrum of Mixed Monomer/Dimer Sample

monomer peaks (ppm)	assignment (monomer)	dimer peaks (ppm)	assignment (dimer)	other peaks (ppm)	assignment
188.7	9	198(split)	9	190.7	carbonyl
163.3, 159.7	4, 2*	165.0, 159.7	4, 2*		
149.4, 147.6	13, 14*	146.4	13, 14*		
137.8	7				
132.7	6				
129.1	11				
128.5	16	126.1**	1,6,11,16*		
124.4	8				
122.9	1				
108.5, 103.0**	5, 15, 12*	107.6, 105.9**	5,15,12*		
101.2	43				
99.3	3				
57.0, 55.0	20, 40*				
		49(split), 40(split)	7,8*		

* these assignments could not be made more specific and are interchangeable.
** not all peaks visible

Conclusions from NMR experiments

There are eighteen peaks in the spectrum of the freshly synthesised sample which can all be assigned to carbons in the monomer molecule. Therefore, this sample is pure monomer. The spectrum of material originally obtained contains thirty-six peaks, half of which are those assigned to the monomer. The other half can be assigned to the dimer molecule which crystallised from a solution of this material. There are no remaining peaks, therefore only one stereoisomer of the dimer is present.

In the spectrum of the solution of monomer/dimer mixture, from which monomer had been removed via crystallisation, the relative intensity of the peaks attributed to the dimer is increased. This shows that monomer and dimer concentrations are not in equilibrium in solution.

The spectrum of the monomer solution was unchanged after the solution was left for several months, showing that dimerisation does not occur in solution at room temperature. Irradiation with light, λ=380 nm, for 90 minutes also failed to bring about a change in the spectrum.

The solid-state spectrum of the freshly synthesised sample contains seventeen peaks, which can all be assigned to the monomer molecule with the aid of solution spectra assignments. The spectrum of the originally obtained solid contains the monomer spectrum peaks. Other peaks can be assigned to the dimer, with the use of solution spectrum assignments, giving a ratio of roughly 50:50 for monomer:dimer concentration.

In solution, the dimer molecule possesses a mirror plane, which is broken in the solid, resulting in the splitting of the peaks. In addition, the mixed monomer/dimer spectrum contained a peak which can be assigned to a carbonyl carbon in a very similar environment to that of the monomer. There are no other peaks which cannot be assigned to either monomer or dimer. The most likely explanation for this third carbonyl peak is the presence of an additional polymorph, or pseudo-polymorph (containing solvent of crystallisation) of the monomer. It is likely that the shift of the carbonyl carbon would be the most sensitive to the difference in environment in the two crystal forms which would explain why the solid state spectrum contains no other additional peaks.

Appendix: Details of structures and their data collection, structure solution and refinement

Data were collected at Oxford on an Enraf-Nonius CAD4-F diffractometer, using Cu-Kα (λ = 1.54180 Å) radiation except for chalcones XII, XXI and XXV, which were collected using Mo-Kα (λ = 0.7107 Å) radiation. Data were collected at Durham on a Rigaku AFC6S diffractometer with a sealed tube source, except for the XXII dimer crystal, for which a Siemens rotating anode was employed, in all cases using Mo-Kα radiation. The data were corrected for Lorentz and polariation effects and absorption.

Structure solution was acheived by direct methods using SHELX-86 (Sheldrick, 1985). The structures were refined at Oxford using the CRYSTALS package (Watkin *et al.*, 1985). At Durham structures were refined or re-refined using SHELXTL-PLUS (Sheldrick, 1985). Full matrix refinement against Fs was performed with all non-hydrogen atoms modelled anisotropically and hydrogen atom positions calculated using a riding model with C-H equal to 0.96 Å and isotropic thermal parameter, u, to 0.08 Å^2. The final refinements were performed using a weighting scheme implemented in SHELXTL, w = $1/(\sigma^2(F)+gF^2)$, with g refined to values between 0.001 and 0.0001.

Table 9. Data collection, refinement and structural parameters.

a	I	XVI	XXI
Space group	←	P*bca*	→
Z	←	8	→
Cell parameters(Å)	*a* = 11.705	*a* = 7.490	*a* = 11.172
	b = 7.522	*b* = 16.782	*b* = 7.806
	c = 32.222	*c* = 26.827	*c* = 28.586
Unique data	2921	3445	1507
Observed data	1775	1845	768
Criterion for observed data	$I > 2\sigma(I)$	$I > 3\sigma(I)$	$I > 2\sigma(I)$
R(Rw) (%)	4.44(5.82)	4.13(5.29)	3.99(3.90)

b	II	X	XI
Space group	$P2_12_12_1$	$P\bar{1}$	$P2_1$
Z	4	2	2
Cell parameters(Å)), (°)	*a* = 5.282	*a* = 8.798, α =107.64	*a* = 6.408
	b = 8.671	*b* = 8.911, β = 102.18	*b* = 10.498,.β = 90.34
	c = 30.670	*c* = 11.298, γ = 105.95	*c* = 11.472
Unique data	2767	2927	1649
Observed data	1829	2573	1429
Criterion for observed data	←	$I > 3\sigma(I)$	→
R(Rw) (%)	5.11(6.09)	5.45(6.58)	3.24(4.09)

c	XXIV_A	XXIV_B	XXV
Space group	$P\bar{1}$	$P\bar{1}$	$P\bar{1}$
Z	4	2	4
Cell parameters(Å), (°)	*a* = 8.345, α = 64.39	*a* = 8.952, α = 69.25	*a* = 7.722, α = 85.82
	b = 13.995, β = 83.46	*b* = 9.696, β = 66.80	*b* = 8.124, β = 85.06
	c = 14.692, γ = 85.10	*c* = 10.418, γ = 68.82	*c* = 29.158, γ = 66.26
Unique data	5751	2927	3011
Observed data	3478	2573	1811
Criterion for observed data	←	$I > 3\sigma(I)$	→
R(Rw) (%)	4.09(5.38)	4.05(5.27)	4.41(5.43)

d	XVII	XIX
Space group	$P2/a$	$P\bar{1}$
Z	4	2
Cell parameters(Å), (°)	*a* = 14.537	*a* = 8.037, α = 77.27
	b = 9.400, β = 107.38	*b* = 8.252 , β = 88.14
	c = 14.894	*c* = 15.421, γ = 66.40
Unique data	3978	3567
Observed data	3176	2531
Criterion for observed data	$I > 3\sigma(I)$	$I > 3\sigma(I)$
R(Rw) (%)	5.27(6.42)	5.03(6.08)

e	VI	XIV
Space group	$P2_1/c$	$P2_1/n$
Z	4	4
Cell parameters(Å), (°)	a = 13.546(4)	a = 7.586(2)
	b = 8.052(4), β = 109.68(2)	b = 16.271(2) , β = 103.95(1)
	c = 15.097(4)	c = 13.858(2)
Unique data	2722	2919
Observed data	1629	1746
Criterion for observed data	F > 4σ(F)	F > 4σ(F)
R(Rw) (%)	4.06(5.79)	4.37(5.72)

f	XXII(monomer)	XXII(dimer)
Structural formula	$C_{18}H_{16}O_5$	$(C_{18}H_{16}O_5)_2$
M	312.31	624.62
Crystal system	Monoclinic	Triclinic
Space group	$P2_1/c$	$P\bar{1}$
Z	4	2
Cell parameters(Å), (°)	a = 3.927(1)	a = 8.837(5), α = 89.90(1)
	b = 15.790(3), β = 91.67(2)	b = 12.729(8), β = 79.87(1)
	c = 24.013(3)	c = 14.452(8), γ = 70.34(2)
Cell volume($Å^3$)	1488.4(6)	1504.1(6)
Unique data	4354	5306
Observed data	1581	3875
Criterion for observed data	F > 4σ(F)	F > 4σ(F)
R(Rw) (%)	5.34(3.91)	5.12(6.87)

References

Bahar, S. (1992) The crystallisation and X-ray structure analysis of chalcones. *Chemistry Part II Thesis, Oxford University.*

Dominguez, X.A., Garcia, S., Williams, H.J., Ortiz, C., Scott, A.I. & Reibenspeig, J.H. (1989) New chalcones from *Mimosa tenuefolia. J. Nat. Prod.* **52**, 864-867.

International Tables for Crystallography, Volume 4. *Kynoch Press, Birmingham.* (1974)

Kitaigorodsky, A.I. (1961) Organic crystal chemistry. *Consultants Bureau, New York.*

Rabinovich, D. & Shakked, Z. (1974) Optical induction in chiral crystals. I. The crystals and molecular structures of 4,4'-dimethylchalcone. *Acta Crystallogr.* **B30**, 2829-2835.

Rabinovich, D. & Schmidt, G.M.J. (1970) Topochemistry. Part XXXIV. Crystal and molecular structure of p'-bromochalcone. *J. Chem. Soc. (B)* 6-10.

Sheldrick, G.M. (1985) Crystallographic Computing 3. *ed. Sheldrick, G.M., Kruger, C. & Goddard, R.J. Oxford University Press.*

SPECINFO Daresbury Laboratory, Warrington.

Van der Sluis, P. (1989) Single crystals and X-ray structure determination. *Thesis, State University Utrecht, the Netherlands.*

Weygand, C. (1929) Systematische untersuchingen zum polymorphismus organischer substanzen. *Annalen der Chemie* **472**, 143-179.

Woodward, R.B. & Hoffmann, R. (1970) The conservation of orbital symmetry. *Verlag Chemie-Academic Press.*

Zheindong, L., Fen. P. & Genbo, S. (1992) *Acta Crystallogr.* **C48**, 714.

Chapter 2:

Crystallisation of proctolin, a pentapeptide Arg-Tyr-Leu-Pro-Thr

Introduction

Peptides as neurotransmitters

Proctolin, a linear pentapeptide, Arg-Tyr-Leu-Pro-Thr, was the first insect neuropeptide to be chemically characterised. It was isolated and identified in the American cockroach, *Periplanta americana* (Brown, 1975) and since found to occur in species of six orders of insects and some other invertebrates such as lobsters. The peptide was found to act as a neurotransmitter in the cockroach and has been shown to play a neuromodulatory role in several species (Konopiñska *et al.*, 1992).

Nerve impulses are transmitted along chains of nerve cells, neurons. An impulse is passed from one cell to the next across the intervening synapse by the agency of a chemical neurotransmitter. These chemicals are released from vesicles in the pre-synaptic membrane, diffuse across the synaptic cleft, which may be about 500 Å in width, and couple with receptors in the post-synaptic membrane. Until the 1960's the amines acetylcholine, norepinephrine, and serotonin were the only well recognised transmitters. Then came the understanding that amino acids including γ-amino butyric acid (GABA), glutamic acid, aspartic acid and glycine could also serve as neurotransmitters. Thus, it was thought that all neurotransmitters belonged to a group of small molecules with molecular weight around 200 Da. Peptide hormones, with chains comprising up to 40 amino acids, are now known to be involved in addition to these classical neurotransmitters (Snyder *et al.*, 1980; Hökfelt *et al.*, 1980).

The functional significance of peptide neurotransmitters is not fully understood. In the mammalian central nervous system classical neurotransmitters were found to be present in only a small population of neurons. It appeared that cells lacking a classical neurotransmitter may produce a peptide. Peptide hormones have also been detected in neurons containing a small-molecule neurotransmitter, suggesting that the two species may have complementary roles. Some examples of pairs of classical and peptide neurotransmitters found in the mammalian central nervous system are listed in Table 1. This coexistence of two types of molecule may be a consequence of evolution. It is possible that peptides were important messengers in lower species, with more efficient, small molecule transmitters developing later (Hökfelt *et al.*, 1987).

The presence of peptides and classical neurotransmitters in the same neuron could have relevance to the understanding of certain disorders of the nervous system. The progress of Alzheimer's disease and senile dementia are accompanied by the degradation of cholinergic neurons responsible for higher brain functions, such as memory and learning. The peptide hormone galanin, coexisting with acetylcholine in these cells, may be implicated in this process (Hökfelt *et al.*, 1987).

Investigation of insect neuropeptides shows that insects possess a hormone system similar to that of vertebrates, with parallels existing between families of neuropeptides with equivalent functions (Konopiñska *et al.*, 1992). Some examples are listed in Table 2.

Table 1. Mammalian classical and peptide neurotransmitter pairs.

Classical transmitter	Peptide	Brain region	Species
dopamine	neurotensin	ventral mesencephalon	rat
acetylcholine	enkephalin	spinal cord	rat
acetylcholine	galanin	basal forebrain	rat, monkey
GABA	somatostatin	thalmus	cat
GABA	somatostatin	cortex, hippocampus	rat, cat, monkey
glycine	neurotensin	retina	turtle

Table 2. Vertebrate and invertebrate hormones with parallel functions.

Invertebrate hormone	Vertebrate hormone
AKH[a] family	glucagon
myotropic hormones: proctolin, Lem-PK[b]	substance P
leucosulfakinins	gastrin, cholecystokinin
melanisation hormones: Bom-MRCH[c]	MSH[d]
allatostatin	somatostatin
bombyxin-II	human insulin, adenocorticotropic mammal hormone

[a]	AKH	adipokinetic hormone
[b]	Lem-PK	leukopyrokinin
[c]	Bom-MRCH	hormone from *Bombyx mori* (silkworm)
[d]	MSH	melanocyte-stimulating hormone

X-ray crystallographic studies on oligopeptides

The discovery of the significant roles played by small peptides as hormones and neurotransmitters has stimulated the study of the structure, activity and relationship between these properties for peptide hormones. Such investigations aim to learn about conformational changes as the molecule moves from solution to receptor and the effect of chemical modifications on structure and activity. In many cases the receptor site has not been located and the only source of information about the receptor is the study of substrate structures and the conformational differences between active and inactive, agonistic and antagonistic analogues.

Linear peptides of chain length 5 to 30 do not crystallise readily. As the chain length increases beyond 2 residues, the number of energetically favourable conformations a molecule may adopt in solution rapidly increases. However, these peptides are not long enough to fold into the stable secondary structure motifs seen in longer peptide chains. The difficulties encountered during X-ray crystallographic analysis of these materials are reflected by the small number of crystal structures for peptide molecules in the 1992 release of the Cambridge Structural Database (Allen *et al.*, 1991), Figure 1. Cyclisation reduces the conformational space accessible to a molecule, so, for a given chain length, the possibility of crystallisation is greater for a cyclic molecule than a linear one. Exceptions possessing some degree of secondary structure include artificial helical molecules, such as poly-leucines, natural helical molecules, such as glucagon (Sasaki *et al.*, 1975) and species which are pseudo-cyclic due to disulphide linkages, such as deamino-oxytocin (Wood *et al.*, 1986) and crambin (Teeter *et al.*, 1993).

Neither methods employed for the crystallisation of small molecules nor techniques used in protein crystallography are guaranteed to succeed in the production of crystals of oligopeptides. This fact is mirrored by the extremely variable nature of the peptide structures which have been elucidated. Some resemble the closely packed crystals of small molecules, while others have more protein crystal character, consisting of aggregates or layers of molecules separated by wide solvent channels. This point is illustrated by a comparison of four crystal structures of leucine and methionine enkephalin, Table 3 (Murray-Rust, 1991). Enkephalins are endogenous pentapeptides which bind to opiate receptors and have potent analgesic activity. The twelve independent molecules in these four structures exhibit two conformations, one extended, the other comprising a β-bend centred on Gly-Gly. Side-chain orientations are different in each molecule.

Table 3. Comparison of Leu and Met enkephalin crystal structures

crystallisation solvent	space group	cell: a.b.c.(Å) α, β, γ (°)	asymmetric unit
Leu-enkephalin			Tyr-Gly-Gly-Phe-Leu

Aqueous Methanol
(Smith & Griffin, 1978;
Blundell *et al.*, 1979)
	A2	a = 31.937 b = 17.084 c = 24.861
		β = 95.54
		Z = 16
			4 nearly identical conformers with a Gly-Gly β-bend, several water molecules

Aqueous DMF
(Karle *et al.*, 1983;
Camerman *et al.*, 1983)
	$P2_1$	a = 18.720 b = 24.732 c = 20.311
		β = 115.86
		Z = 8
			4 extended conformers forming an antiparallel β-sheet, 8 water and 8 DMF molecules plus some disordered solvent.

Ethanol
(Griffin *et al.*, 1986)
	$P2_1$	a = 11.549 b = 15.587 c = 16.673
		β = 92.19
		Z = 4
			dimers of molecules in extended conformation, related by a pseudo 2-fold axis, 1 water molecule per dimer

Met-enkephalin			Tyr-Gly-Gly-Phe-Met

1% aqueous pyridine +
0.05% acetic acid
(Griffin *et al.*, 1986)
	$P2_1$	a = 11.607 b = 17.987 c = 16.519
		β = 91.24
		Z = 4
			similar to Leu-enkephalin from ethanol, but with 10.6 water molecules per dimer.

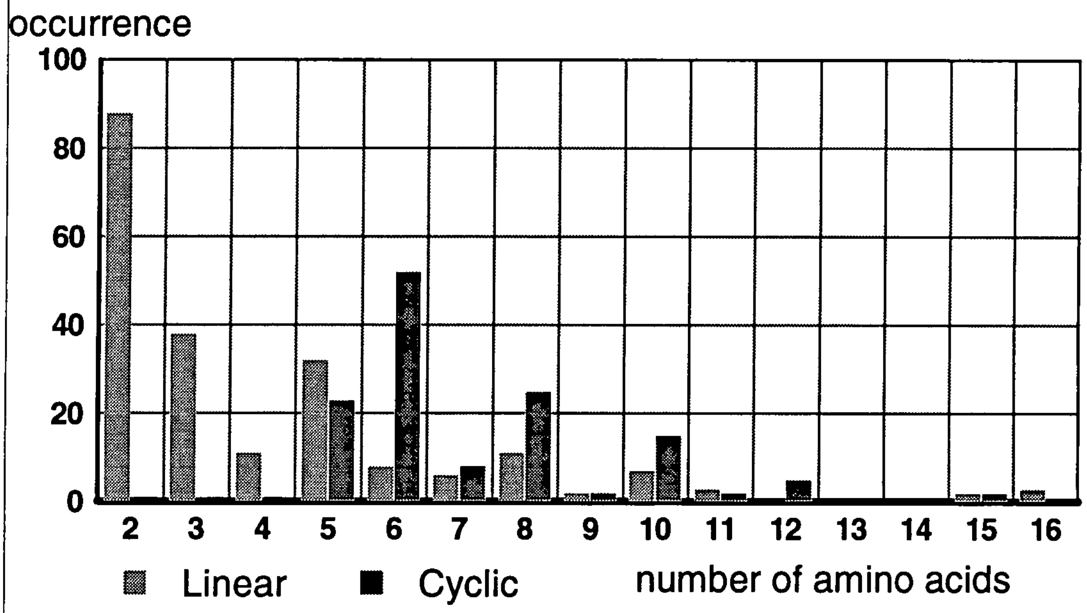


Figure 1. Polypeptides in the CSD

About proctolin

The long-term goal of this project was the development of a new generation of non-cholinergic insecticides. The utilisation of peptide hormones as pesticides is unfeasible. The cost of industrial scale synthesis of these molecules would be prohibitive, but even if this were overcome, externally applied peptides would undergo degradation in the environment or the digestive tract of the target species. The development of small molecule analogues with chemistry tailored to allow their external application as well as an enhancement of agonistic activity is necessary. The aim of this project was to deduce the three-dimensional structure of the active conformation of proctolin to shed light on the nature of the interaction between proctolin and its receptor, providing input for molecular modelling calculations to assist in the design of small molecule proctolin analogues.

Structure-activity relationships for proctolin have been investigated. The activity of analogues in which one or more residues have been altered has been determined. The replacement of residues 3, 4 and 5 leads to inactivity. The analogues Lys-Tyr-Leu-Pro-Thr and Arg-Arg-Leu-Pro-Thr do possess activity. Oxymethyl Tyr possesses enhanced activity and replacement of Tyr by a para-substituted phenylalanine does not lead to complete loss in activity. It has been concluded that a positively charged residue at (1) and a para-substituted aromatic ring at (2) play a crucial role. It has also been observed that the cyclic analogue displays enhanced agonistic activity. This could be because its rigidity makes it difficult to expel it from the receptor, although this does not necessarily imply that the structure of the cyclic form resembles the active conformation of the linear peptide (Temussi *et al.*, 1989).

Search for crystallisation conditions for proctolin

undertaken at Birkbeck College Crystallography Department as part of this PhD

Aim

The physical and chemical nature of the material was explored, to provide indications of possible conditions for crystallisation. The intention was to develop and refine a crystallisation method, eventually producing crystals of X-ray diffraction quality. Repurification of old material and automated peptide synthesis were undertaken to supply sufficient peptide for this series of trials.

Material

Composition:	1 M proctolin + 3.5 M water + 0.5 M acetate
Anhydrous molecular weight (free base):	648.8
Molecular weight (with water and acetate):	746.3
Appearance:	white solids
HPLC purity:	99%

Purification and analysis by HPLC

The purification of peptides becomes simpler as the chain length decreases and the yields increase. The retrieval of material from old experiments for reuse is feasible for a five residue peptide.

Rinsings of material from old experiments were pooled and the solvent removed using a vacuum centrifuge. An aqueous solution of the material, of concentration approximately 1 mg/ml, was made up. Concentration was measured from the absorbance at 280 nm;

absorption coefficient of 1 mg/ml solution at 280 nm $= 1300\ N_{Tyr} / M_{molecule}\ (au^{-1})$

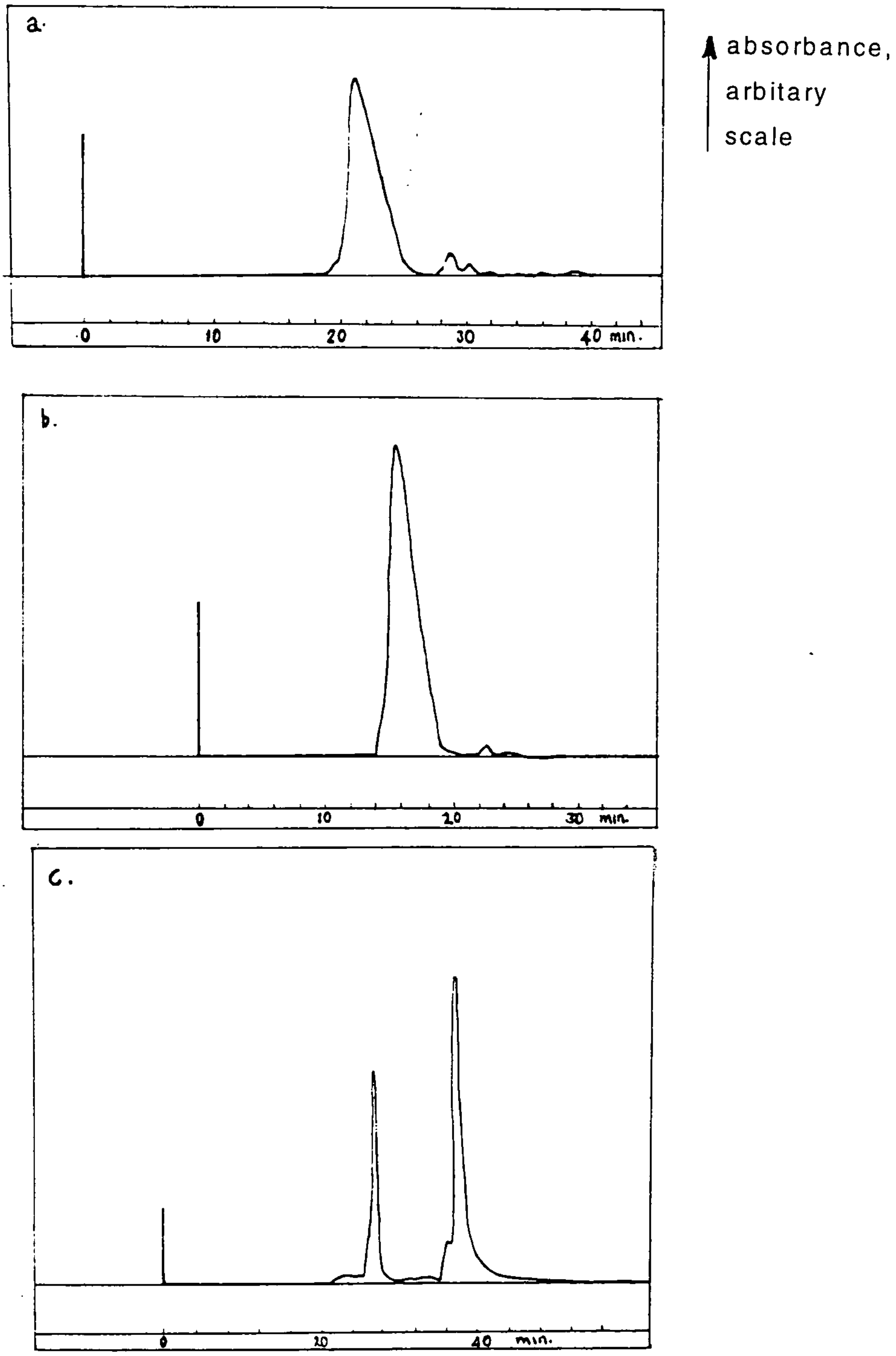
$= \quad 1300 \times 1 / 648 \quad = \ 2$

Analysis of the material and separation of the peptide from the accumulated impurities was carried out by HPLC, Figure 2. Each HPLC run gave a large major peak, representing the elution of proctolin. The position of this peak varied between samples over a concentration range of 18.5 - 26.5 % B (as designated in Figure 2). Some samples also had minor peaks, eluted at higher B concentration. From this it can be concluded that the separation of proctolin from impurities may require more steps if it is to yield material of crystallisable quality.

The 'proctolin' fractions were collected together, reduced to a volume of 1 ml on a rotary evaporator and then freeze-dried. This material differs from that supplied by Sigma, in that the counter-ion is tri-fluoro acetate, rather than acetate. The material showed significant differences in its behaviour in comparison to the original material. The change in counter-ion is one contributory factor, in addition to the generally high level of impurities. In conclusion, retrieval of material from old crystallisation experiments with a useful yield is probably not possible.

Figure 2,

HPLC spectra of proctolin samples retrieved from old experiments, run on a Brownlee C18 300 Å 7.5 x 250 mm column, using a reversed phase program; solvent A: 0.1 % aqueous trifluoroacetic acid, solvent B: 0.1 % trifluoroacetic acid in 3:1 acetonitrile:water, gradient: 0 → 50 % B over 50 minutes, flow rate: 2 ml/minute, detector: 280 nm absorbance. *a*: 1 division ≡ 2 minutes, major peak at 21.25 % B, *b*: 1 division ≡ 2 minutes, major peak at 15.5 % B, *c*: 1 division ≡ 4 minutes, major peaks at 26.5 %, 37.5 % B - the second peak is due to the presence of a picrate complex.



Solid phase peptide synthesis

Introduction

The basis of this method is that the growing peptide chain remains attached to an insoluble support, consisting of beads of cross-linked polystyrene resin, from the first to the final step of the synthesis. Between each reactive step, the supporting resin and attached peptide are thoroughly washed, removing impurities and reaction by-products while keeping the loss of target peptide to a minimum. The dry beads have diameter 40 - 100 microns, swelling one-hundredfold on the addition of solvent, typically dichloromethane or dimethyl formamide. Macroscopically, the resin appears insoluble, while on a molecular level it is fully solvated, allowing reactants to approach the anchored peptide chain.

The C-terminal amino-acid is attached to the resin by its α-carboxyl group and thus the peptide is assembled from C to N terminal, by consecutive coupling reactions in which amino-acids are added, one at a time, to the chain. Amino-acids are introduced into the reaction vessel with N terminal and side chains protected and C-terminal activated for the coupling reaction. Peptide bond formation is followed by the deprotection of the new N terminal. The activation, coupling and decoupling cycle is repeated for successive amino-acid additions. Finally, the completed peptide chain is cleaved from the resin and the side chains are deprotected. The whole procedure is run by a program with parameters set before the commencement of the synthesis and it is completely automated. Thus, the time and skill required to synthesise a peptide in this manner are a small fraction of what is required for a solution phase synthesis.

In solid phase peptide synthesis, the N-terminal amine is typically protected by the carboxy-amide with either a Boc (t-butoxyloxycarbonyl) or F-moc (9-fluorenylmethoxycarbonyl) group. The F-moc group is removed by base hydrolysis, using a weak base such as a secondary amine, commonly piperidine. Removal of Boc is accomplished by acid hydrolysis, using a weak acid, often tri-fluoro acetic acid. Side chain protecting groups must be stable under N-terminal deprotecting conditions, so complementary protection is required. If a synthesis employs F-moc protection for the N terminal, side chain protection should be acid labile and base stable.

Activation of the attacking C terminal α-carboxyl group allows efficient peptide bond formation, without rearrangement. N,N'-dicyclohexylurea (DCC) is the activating agent commonly employed. The α-carboxyl of the amino acid attacks DCC, forming an O-acylisourea. This derivative may be added directly to the reaction vessel. In excess of amino acid, the O-acylisourea is attacked again, giving the anhydride, an alternative activated species for introduction to the resin. In addition, HOBt is often added to the reaction vessel. HOBt attacks the O-acylisourea, forming an active ester, which undergoes highly efficient coupling with the free peptide N-terminal.

Recently, activation using 2-(1 H-benzotriazol-1-yl)-1,1,3,3-tetramethyluronium hexafluorophosphate (HBTU) has been employed. HBTU is dissolved in a solution of HOBt in DMF and the amino acid added, giving the HOBt active ester. The speed of this activation step significantly reduces the total synthesis time.

Automated solid-phase peptide synthesis of proctolin

The synthesis of proctolin presents one potential problem. The proline residue is prone to attack by the unprotected N-terminal of the succeeding residue, in this case leucine, particularly when it is near the C-terminal of the chain and therefore close to the resin support. This attack causes rearrangement and removal of the peptide from the resin. Reduction of the time taken for steps in which the leucine N-terminal is free should help minimise this effect.

The synthesis employed 'Fastmoc' chemistry, with p-hydroxymethylphenoxymethylpolystyrene (HMP) resin, HBTU.HOBt activation and F-moc protected amino acids: F-moc-Arg(Mtr), F-moc-Tyr(tBu), F-moc-Leu, F-moc-Pro and F-moc-Thr(tBu).

The initial step was the substitution of the N-protected threonine on to the resin. A Thr-resin sample was removed from the reaction vessel for determination of the substitution efficiency. Any unsubstituted sites on the resin were then capped by reaction with benzylic anhydride. Activation, coupling and deprotection steps followed as the four remaining amino acids were successively added. Two coupling cycles were performed for both arginine and tyrosine, since these are known to have poorer coupling efficiencies. Following each coupling cycle, a small sample of the resin was removed for analysis of the coupling efficiency.

Substitution determination

The substitution in mmol/gram of an F-moc protected amino-acid on to the resin may be determined by complexation of the F-moc removed from a preweighed sample of the substituted resin with piperidine. The concentration of the F-moc-piperidine complex is assayed by measurement of the 301 nm absorbance.

The resin sample was dried, then weighed. 0.5 ml 20% piperidine in DMF was added. The solution was shaken for 15 minutes, then made up to a volume of 10 ml. Absorbance was measured.

Sample mass:	1.3 mg			
Absorbance:	0.674			
Substitution (mmol/gram)	=	A_{301} x V(ml) / 7800 m(g)	=	0.67 mmol/gram
Total mass of resin:	278 mg			
Amino-acid substituted on the resin:		278 x 0.67	=	0.186 mmol

Quantitative Ninhydrin monitoring for determination of coupling efficiencies

Ninhydrin monitoring measures the concentration of free amine in the resin sample withdrawn from the reaction vessel following a coupling step. The most recently added amino acid remains F-moc protected at this stage, so the concentration of free amine represents the proportion of vacant sites at which coupling failed.

Samples were removed from the reaction vessel into preweighed tubes, to which 2-3 ml of methanol and 2-3 drops of acetic acid had been added. After the conclusion of the synthesis, the methanol was pipetted off, the resin was rinsed, then dried and the weight determined. A reagent mixture was added to each tube, including a control empty tube, as follows: 75 µl phenol in ethanol + 100µl KCN in pyridine + 75 µl ninhydrin in ethanol. 5 minutes incubation at 100° followed, then the volume was immediately made up to 5 ml with 60% ethanol. The tubes were shaken to ensure thorough mixing. Absorbance at 570 nm was measured, against a 60 % ethanol blank:

Free amine concentration (μmol/g):10^6 (A_{570} x dilution(ml)) / (extinction coefficient x m_{sample} (mg))

dilution:	5 ml
extinction coefficient:	15 000 $M^{-1}cm^{-1}$
mass resin before synthesis;	0.278 g
substitution of first residue on to resin:	0.67 mmol / gram
results	Table 4

Table 4. Coupling efficiency of each stage of the synthesis

residue	coupling efficiency (%)	predicted amount of peptide on resin (mmol)
Thr	99.79	0.185
Pro	99.85	0.184
Leu	99.57	0.182
Tyr	99.68	0.181
Arg	99.58	0.180
total	98.48	0.180

Cleavage

Cleavage of the peptide from the resin and simultaneous side chain deprotection were accomplished by acid hydrolysis, using trifluoroacetic acid. Acid hydrolysis of O-t-Bu protecting groups produces t-butyl cations and t-butyl trifluoroacetate, which may attack the deprotected peptide. Cooling reduces the rate of these reactions, but also the cleavage and deprotection rates. A better solution is the introduction of scavengers to remove the cations. 1,2-ethane dithiol is an efficient scavenger for t-butyl trifluoroacetate. Addition of a second scavenger, such as anisole, phenol, ethyl methyl sulphide, thioanisole, 2-mercaptoethanol, thiophenol, tryptophan or methionine, is necessary for the complete suppression of the alkylating reactions.

The following cleavage reagent was prepared: 0.75 g crystalline phenol + 0.25 ml ethanol + 0.5 ml water + 10 ml TFA. This was cooled in ice, then added to the dry peptide-resin in Eppendorf tubes, which were then sealed. The mixture was allowed to warm to room temperature and shaken at regular intervals over a three hour period. The solution was filtered from the resin through glass wool. The peptide was precipitated in excess diethyl ether. The precipitant was washed, dried and weighed.

Yield

The total yield of dry product, before purification, was 106.6 mg.

Crystallisation

Introduction : finding crystallising conditions for peptides

This account summarises discussions with Marek Brzozowski at York and Steve Wood at Birkbeck.

When it comes to crystallisation, polypeptides may be divided into three groups. Small peptides, with 1 to 4 amino acids, medium, with 5 to around 30 and large, the remainder. Small peptides behave as small organic molecules and methods of crystallisation developed for organic molecules are applicable. Crystallisation from an evaporating solution is possible. It is best to make up the peptide solution, then allow around three days equilibration in a sealed vessel before evaporation commences. The interactions in solution are complex and the attainment of a minimum energy conformation may take time. Ethanol is a useful solvent, other possibilities include methanol, DMSO, dioxane and toluene. A mixture of solvents may improve results, for instance 2:1 Ethanol:DMSO. A polar / non-polar solvent mixture, such as 2:1 ethanol: cyclohexane may increase chances of crystallisation. It is possible that inclusion of both types of solvent within the crystal could help stabilise the polar and apolar environments around different parts of the peptide molecule.

Medium peptides, with their increasing number of accessible conformations and complexity of chemical properties, represent more of a challenge. Methods successful in crystallising smaller chains may be attempted. More precision in the reproduction of conditions and stricter purity of solutions are often required. Crystals are likely to be stable in a narrow range of conditions only. Long periods of equilibration followed by slow changes in the environment are needed and, on the appearance of crystals, rapid mounting and data collection are advisable.

The search for crystallisation conditions initially entails a determination of the solubility of the material in water and organic solvents. Ways of manipulating this solubility must be explored. This is equivalent to mapping out the border between solution and precipitation in a multidimensional phase space. It involves experimenting with solvent mixtures. The effect of changing pH should be investigated. The presence of additives and change in temperature are other parameters that need to be considered. This approach is summarised in the flowchart, Figure 3.

Charged groups present particular problems. One approach could be the crystallisation of material from which protecting groups were not removed following synthesis. However the resulting crystal structure might bear only a poor resemblance to the native form. Addition of a counter ion resulting in precipitation of a complex is a less extreme possibility, for example, flavine was co-crystallised with oxytocin. The counter ion may stabilise a crystal lattice by acting as a bridge between regions of like charge in the peptide molecules.

The division between medium and large peptides is determined by the point at which the peptide chain has sufficient length to fold into globular structures containing stable secondary structural units. In the presence of favourable intramolecular interactions, such as disulphide linkages, the chain length required for the existence of secondary or tertiary structure is reduced.

In protein crystals, loosely packed molecules are interspersed with solvent channels. The solvent content is typically around 50 %, see Chapter 5B. The crystals are often fragile and stable only under a solvent atmosphere. The protein molecule is a complex chemical system, thus the growth of crystals is mediated by a complex and precise interplay of external factors. The amount of material available for experiment may be small, a few tens of mgs, or less. Systematic methods for the crystallisation of proteins have been developed, utilising comparatively tiny quantities of material (Ducruix & Giege, 1992). Crystal growth by vapour diffusion from hanging drops is the usual method employed, Figure 4. A widely used procedure is the performance of an initial screening for precipitation inducing conditions using a set of 50 precipitant, buffer and salt combinations suggested by a factorial analysis (Ducruix & Giege, 1992; Jancarick & Kim, 1991). Because a comparatively small amount of material is required to set up such a screening experiment, it is justifiable for a medium length peptide, although the chances of finding appropriate conditions are reduced.

Figure 3. The search for crystallisation conditions for a peptide of medium length

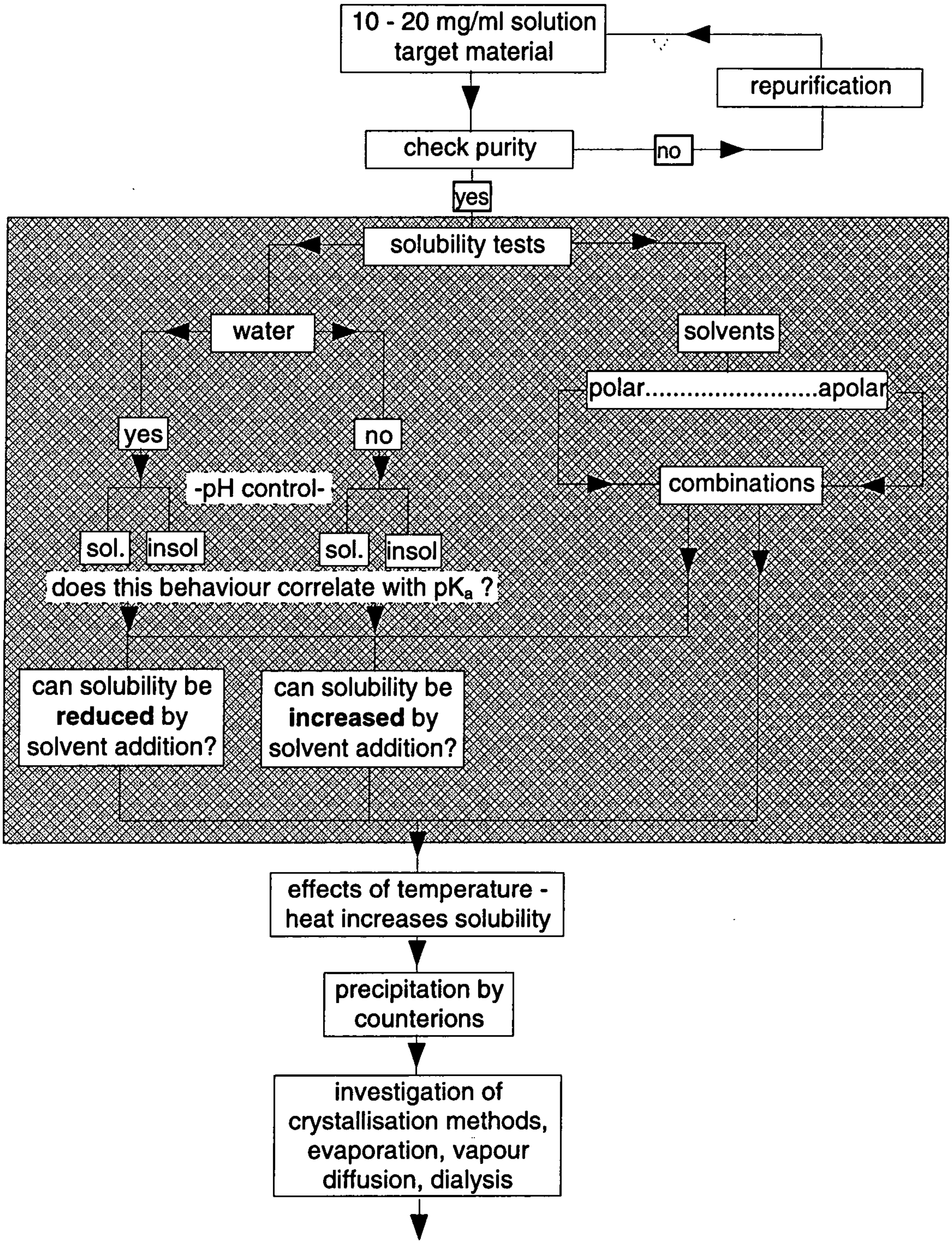
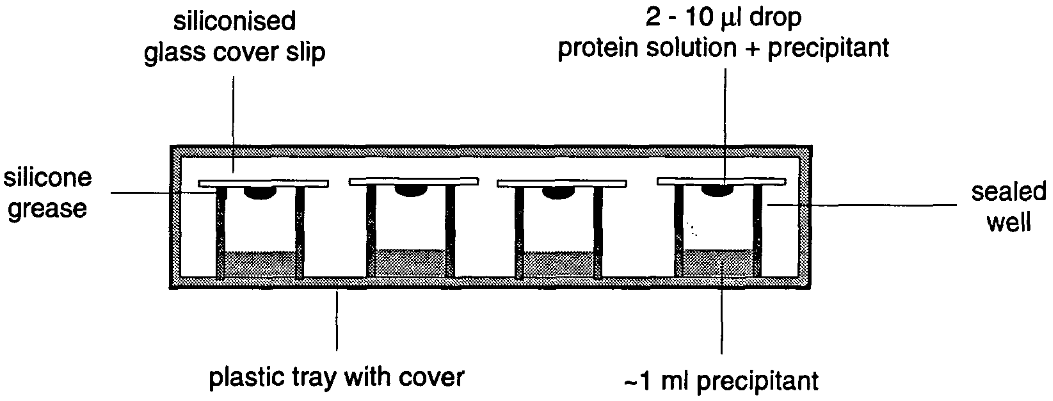

10 - 20 mg/ml solution target material
repurification
check purity
no
yes
solubility tests
water
solvents
polar.........................apolar
yes
no
-pH control-
combinations
sol.
insol
sol.
insol
does this behaviour correlate with pK$_a$?
can solubility be **reduced** by solvent addition?
can solubility be **increased** by solvent addition?
effects of temperature - heat increases solubility
precipitation by counterions
investigation of crystallisation methods, evaporation, vapour diffusion, dialysis

Figure 4. hanging drop vapour diffusion method of crystallisation



Solubility tests

Proctolin is extremely soluble in water, methanol, ethanol and iso-propanol, giving 40 mg/ml or higher solution concentrations. Solubility across a range of pHs was tested, using a series of phosphate/citrate buffer solutions. The material was soluble at pHs 2.2 - 8.0 in water, 1:1 ethanol:water and ethanol. No precipitation occurred when the solutions were cooled to 4 °C.

A 20 mg/ml aqueous solution was made up and drops of organic solvent; DMSO, DMF, dioxane, acetonitrile, butan-1-ol, acetone and ethyl acetate, were added. Precipitation was observed on addition of dioxane. The following organic solvents were added to alcoholic proctolin solutions; diethyl ether, dichloromethane, chloroform, toluene, hexane, cyclohexane and carbon tetrachloride. Precipitation from a 20 mg/ml ethanol solution was seen on addition of ether, chloroform, and dichloromethane, from a 20 mg/ml methanol solution on addition of ether and toluene. The greatest degree of precipitation was observed on addition of diethyl ether to an ethanol solution.

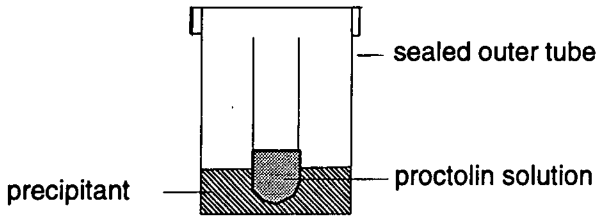
Complexation of the arginine residue could lead to precipitation. A saturated aqueous solution of picric acid was added to aqueous and alcoholic proctolin solutions buffered at pH 4. Precipitation was observed, to the greatest extent in a 1:1 ethanol:water solution. Evaporation of these solutions resulted in the growth of picrate crystals. Precipitation was also observed in a 10 mg/ml proctolin solution in 50% acetic acid.

Crystallisation tests

Concentrated solutions of proctolin were put into siliconised test tubes, sealed and left for several days. No encouraging results were obtained by this batch method.

With indications of possible solvent/precipitant pairs from the solubility test results, vapour diffusion experiments were set up, using the apparatus shown in Figure 5. Ethanol and methanol solutions of proctolin were used, with diethyl ether, chloroform and dichloromethane as precipitants. Precipitation was observed after 1-2 days. The precipitate was mainly in the form of oily droplets. Tests using ether as precipitant contained some material which appeared more crystalline.

Figure 5. Apparatus for crystallisation by 'batch' vapour diffusion



The combination of an ethanol solution with ether as a precipitant was the most promising approach, from the trials up to this point. A recipe for crystallisation was developed as follows:

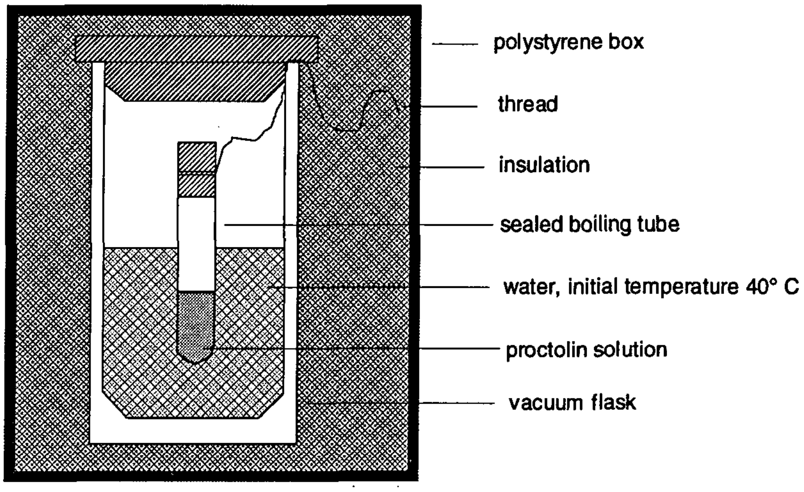
Ether was added dropwise to a test tube containing a 40 mg/ml ethanolic solution of proctolin. Precipitation was observed to form and then redissolve. Addition of ether was continued until the point at which the precipitate remained. The tube was stoppered and placed in a water bath at 40° C for ten to fifteen minutes, after which time the precipitate had redissolved. The tube was allowed to cool slowly to room temperature, then placed in a cold room, at 4° C.

Crystals grew in three to five days, forming clusters of extremely thin plates. It was possible to separate individual crystals from the clusters. The length and width of the largest were 0.35 mm x 0.10 mm but their thickness was too small to measure. A crystal was mounted in a capillary and X-ray photographs were taken. No diffraction was observed.

Attempts were made to improve crystal quality by altering the concentrations of solvent and precipitant. The above recipe produced the best results. Ethanol and ethanol/ether solutions of proctolin were seeded with crystals obtained by the above method, but no improvement was observed.

One crystallisation experiment was set up with essentially the conditions described above, but with increased size and greater control over the rate of change of temperature. 25 mgs of proctolin were dissolved in 0.625 ml of ethanol. The solution was centrifuged to remove any solid contaminant, then pipetted into a large test tube. Ether was titrated into the solution until the point of stable precipitation was reached. The tube was sealed and heated in a water bath at 40° C for fifteen minutes, when the precipitate had redissolved. The tube was then placed inside a slow-cooling apparatus, Figure 6, which was sealed and left in the cold room, at 4° C, for three weeks. Crystals had grown of a very similar shape to those previously produced. The maximum length and width had increased considerably, to around 5 mm x 0.3 mm, but the crystals remained very thin.

Figure 6. Slow cooling apparatus



Crystallisation was also attempted using the set of 50 conditions specified in the Hampton Research Crystal Screen (Jancarik & Kim, 1992) which are commonly used for initial trials in protein crystallography. A 20 mg/ml aqueous solution of proctolin was used. Two crystallisation methods were employed, sitting-drop vapour diffusion, using Linbro plates, and crystallisation under an oil film. Absolutely no precipitation was observed.

Conclusions

The first problem encountered during the attempt to crystallise proctolin was its extreme solubility in aqueous and alcoholic solutions. Precipitation could only be induced from very concentrated solutions, therefore large amounts of material were required. Attempts to re-use old material and synthesise new did not succeed in producing material with the same properties as that supplied by Sigma. The newly synthesised proctolin requires ion exchange and further purification steps. Crystallisation was achieved, but the crystals were not suitable for diffraction studies. The extreme thinness made it difficult to harvest and mount the crystals from solution in the test tube in which they had grown. They were only stable under ethanol vapour and possibly the presence of ether is also required. No diffraction was observed, possibly due to mounting problems, or the meagre volume resulting from the small third dimension.

A pentapeptide such as proctolin fluctuates in solution between many energetically similar conformations with low potential barriers between them. A unique bioactive conformation is adopted at the receptor site as a consequence of the peculiar environment. There is no obvious relationship between the bioactive and lowest energy solution conformations. Solution NMR studies provide information about conformations stabilised by an aqueous environment, while most active sites are known to be hydrophobic cavities. Crystallisation introduces yet another environment, the molecular conformation within a crystal being mediated largely by lattice forces. The presence of additives and counter ions in a crystal introduce further complexity. A single structure determination in solution or solid phase is thus of limited relevance when the goal is an understanding of the activity of the molecule. Putting together several pieces of structural information, coupled with modelling studies, may allow some speculations to be made.

References

Allen, F.H., Davies, J.E., Galloy, J.J., Johnson, O., Kennard, O., Macrae, C.F., Mitchell, G.F., Smith, J.M. & Watson, D.G. (1991) *J. Chem. Inf. Comput. Sci.* **31**, 187-204

Blundell, T.L., Hearn, L., Tickle, I.J., Palmer, R.A., Morgan, B.A., Smith, G.D. & Griffin, J.F. (1979) Crystal structure of [Leu^5] enkephalin. *Science* **205**, 220.

Bodansky, M. (1984) Principles of peptide synthesis. *Springer-Verlag, New York.*

Brown, B.E. (1975) Proctolin: a peptide transmitter candidate in insects. *Life Sci.* **17**, 1241-1252.

Camerman, A., Mastropaolo, D., Karle, I.L., Karle, J., & Camerman, N. (1983) Crystal structure of leucine enkephalin. *Nature* **306**, 447-450.

Ducruix, A., & Giege, R. - ed. (1992) Crystallisation of nucleic acids and proteins: a practical approach. *Oxford University Press.*

Griffin, J.F., Langs, D.A., Smith, G.D., Blundell, T.L., Tickle, I.J. & Bedarkar, S. (1986) The crystal structure of [Met^5] enkephalin and a third form of [Leu^5] enkephalin: observations of a novel pleated β-sheet. *Proc. Natl. Acad. Sci. USA* **83**, 3272-3276.

Hökfelt, T., Johansson, O., Ljungdahl, A., Lundberg, J.M. & Schultzberg, M. (1980) Peptidergic neurones. *Nature* **284**, 515-520.

Hökfelt, T., Millhorn, D., Seroogy, K., Tsuruo, Y., Ceccatelli, S., Lindh, B., Meister, B., Melander, T. & Shalling, M. (1987) Coexistence of peptides with classical neurotransmitters. *Experimentia* **43**, 768-776.

Janckarik, J. & Kim, S.H. (1991) Sparse matrix sampling: a screening method for crystallization of proteins. *J. Appl. Cryst.* **24**, 409-411.

Karle, I.L., Karle, J., Mastropaolo, D., Camerman, A. & Camerman, N. (1983) [Leu^5] enkephalin: four cocrystallizing conformers with extended backbones that form an antiparallel β-sheet. *Acta Crystallogr.* **B39**, 625-637.

Konopiñska, D., Rosiñski, G. & Sobótka, W. (1992) Insect peptide hormones, an overview of the present literature. *Int. J. Peptide Protein Res.* **39**, 1 - 11.

Murray-Rust, J. (1991) Crystallography in drug design. *Chapter 3, Peptide Pharmaceuticals, editor D.J. Ward, Open University Press, Milton Keynes.*

Sasaki, K., Dockerill, S., Adamiak, D.A, Tickle, I.J. & Blundell, T. (1975) X-ray analysis of glucagon and its relationship to receptor binding. *Nature* **257**, 751-7.

Smith, G.D. & Griffin, J.F. (1978) Conformation of [Leu^5] enkephalin from X-ray diffraction: features important for recognition at opiate receptor. *Science* **199**, 1214-1216.

Snyder, S.H. (1980) Brain peptides as neurotransmitters. *Science* **209**, 976-982.

Teeter, M.M., Roe, S.M. & Heo, N.H. (1993) Atomic resolution (0.83 Å) crystal structure of the hydrophobic protein crambin at 130 K. *J. Mol. Biol.* **230**, 292-311.

Temussi, P.A., Picone, D., Castiglione-Morelli, M.A., Motta, A. & Tancredi, T. (1989) Bioactive conformation of linear peptides in solution: an elusive goal? *Biopolymers* **28**, 91-107.

Wood, S.P., Tickle, I.J., Traharne, A.M., Pitts, J.E., Mascarenhas, Y., Li, J.Y., Husain, J., Cooper, S., Blundell, T.L., Hruby, V.J., Buku, A., Fischman, A.J. & Wyssbrod, H.R. (1986) Crystal structure analysis of diamino-oxytocin. *Science* **232**, 633-636.

Operation manual for automatic peptide synthesizer. (1993) *Applied Biosystems, Inc.*

Chapter 3

Structural studies on glucosamine 6-phosphate synthase

Abbreviations and definitions used in Chapter 3

R_{merge}	$\Sigma_h \Sigma_{i=1}^{N} \mid <I(\mathbf{h})> - I(\mathbf{h})_i \mid / \Sigma_h \Sigma_{i=1}^{N} I(\mathbf{h})_i$ (%), $I(\mathbf{h})_i$ = ith of N measurements of reflection **h**
R_{native}	$\Sigma_h \mid <F_{PH}(\mathbf{h})> - <F_P(\mathbf{h})> \mid / \Sigma_h <F_P(\mathbf{h})>$ (%), $F_P(\mathbf{h})$ = amplitude of reflection **h** in native, $F_{PH}(\mathbf{h})$ = amplitude of reflection **h** in 'derivative'
Patterson synthesis	$P(x,y,z) = 1/V \, \Sigma_h \Sigma_k \Sigma_l \, \lvert F_{hkl} \rvert^2 \cos 2\pi (hx + ky + lz)$
isomorphous Patterson	Patterson synthesis with coefficients $(\Delta_{iso})^2$, $\Delta_{iso} = \lvert F_{PH} - F_P \rvert$
anomalous Patterson	Patterson synthesis with coefficients $(\Delta_{ano})^2$, $\Delta_{ano} = \lvert F_{PH}(+) - F_{PH}(-) \rvert$
k_{emp}	$2 \sqrt{(< F_{PH} - F_P >^2 / < F_{PH}(+) - F_{PH}(-) >^2)}$ for acentric reflections
GAT domain	glutamine amide transfer domain of amidotransferase enzyme
synthase / synthetase domain	amidotransferase domain responsible for amination of a substrate. Synthetase, if a cofactor is involved, synthase otherwise
Ntn amidotransferase	amidotransferase subfamily defined by N terminal catalytic activity.
Triad amidotransferase	amidotransferase subfamily defined by catalytic triad in the GAT domain.
GLMS	glucosamine 6-phosphate synthase
PURF	PRPP amidotransferase
PRPP	phosphoribosyl pyrophosphate
DON	6-diazo-5-oxonorleucine, an inhibitor of Ntn amidotransferases
FMDP	fumaroyl di-propionic acid methyl ester, inhibitor of Ntn amidotransferases

Introduction

Glucosamine-6-phosphate is a member of the glutamine dependent amidotransferase class of enzymes. Glutamine constitutes a major source of nitrogen for biosynthetic processes. These enzymes are responsible for hydrolysis of glutamine and subsequent amination, as depicted in (1), of a range of substrates, including amino acids, nucleotides, sugars, coenzymes and antibiotics. The glutamine hydrolysis (2) and substrate amination (3) reactions occur at separate sites, on different domains known as glutamine amide transfer, GAT, and synthetase, if the reaction involves a cofactor, or synthase, where no cofactor is required. Although separated domains often exhibit limited activity, they function co-operatively, resulting in an efficient coupling of the two halves of the reaction. In some cases the two domains constitute distinct subunits, in others they are present on the same chain. The mechanism by which amide is passed between domains is unclear. There is evidence that the product of glutamine hydrolysis does not equilibrate with native ammonia. However, most amidotransferases can utilise free ammonia in the absence of glutamine. The glutaminase activity is dependent on an active site cysteine, since addition of an irreversibly binding glutamine analogue, such as 6-diazo-5-oxonorleucine, DON, which complexes with the cysteine and mutation of the cysteine, both result in inactivity.

$$\text{Gln} + \text{R} \rightarrow \text{R-NH}_3^+ + \text{Glu}^- \qquad (1)$$

$$\text{Gln} + \text{OH}^- \rightarrow \text{Glu}^- + \text{‘NH}_3\text{’} \qquad (2)$$

$$\text{R} + \text{‘NH}_3\text{’} + \text{H}^+ \rightarrow \text{R-NH}_3^+ \qquad (3)$$

The enzymes can be divided into two subfamilies on the basis of conserved GAT domains, the synthase domains of each enzyme being generally unrelated. The members of one family are characterised by the possession of a Cys-His-Glu catalytic triad and are designated triad amidotransferases. The other type, distinguished by a catalytic N-terminal Cys, forms the N-terminal nucleophile, or Ntn subfamily. Examples of each are listed in Table 1. The Ntn subfamily are members of a wider group of Ntn amidohydrolases, including penicillin acylase (Brannigan *et al.*, 1995). Triad amidotransferases have a GAT domain of around 190 residues, which may or may not be a distinct subunit. The catalytic cysteine lies in the central conserved region of the chain, at around position 90. Ntn enzymes all have fused GAT and synthase domains, the GAT constituting the N-terminal domain and comprising 148-202 residues. The active cysteine is the N-terminal residue of the mature enzyme.

Sixteen amidotransferases are known, four of which are members of the Ntn family, as listed in Table 1. The X-ray crystal structure of one of these, PRPP amidotransferase or PURF, has been determined (Smith *et al.*, 1994). The X-ray crystal structure of one member of the Triad amidotransferase family, GMP synthetase, has also been solved (Tesmer *et al.*, 1996).

Table 1. Examples of glutamine dependent amidotransferases

enzyme	*pathway*	*reaction catalysed*	*class*
anthranilate synthase	Tryptophan	chorismate + Gln $\xrightarrow{Mg^{2+}}$ anthranilate + pyruvate + Glu	Triad
carbamoyl-P synthetase	Arginine, UTP, CTP	HCO_3 + 2ATP + Gln → $NH_2CO_2PO_3^{2-}$ + 2ADP + P*i* + Glu	Triad
CTP synthetase	CTP	UTP + ATP + Gln $\xrightarrow{Mg^{2+}}$ CTP + ADP + P*i* + Glu	Triad
GMP synthetase	GMP	XMP + ATP + H_2O + Gln → GMP + AMP + PP*i* +Glu	Triad
asparagine synthetase	Asparagine	Asp + ATP + Gln $\xrightarrow{Mg^{2+}}$ Asn + AMP + PP*i* Glu	Ntn
glucosamine 6-P synthase	Glucosamine 6-P	fructose-6-P + Gln → glucosamine 6-P + Glu	Ntn
glutamate synthase	Glutamate	α-ketoglutarate + NADPH + H^+ + Gln → $NADP^+$ + 2Glu	Ntn
glutamine PRPP amidotransferase	AMP, GMP	PRPP + H_2O + Gln → phosphribosylamine + P*i* + Glu	Ntn

Glucosamine 6-phosphate synthase, GLMS, is an Ntn amidotransferase. It catalyses the first step in the hexosamine biosynthetic pathway (4). The product, glucosamine 6-phosphate is the precursor of UDP-N-acetyl glucosamine from which all amino sugar containing macromolecules are derived in both prokaryotic and eukaryotic cells. The inhibition of this enzyme could consequently play a role in the treatment of bacterial and fungal infection, certain types of diabetes and some cancers.

$$\text{fructose-6-P} + \text{Gln} \rightarrow \text{glucosamine 6-P} + \text{Glu} \qquad (4)$$

$$\text{Gln} + H_20 \rightarrow \text{Glu} + \text{‘}NH_3\text{’}$$

GAT domain

$$\text{fructose-6-phosphate} + \text{‘}NH_3\text{’} \rightarrow \text{glucosamine-6-phosphate}$$

synthase domain

The mechanism of glutamine hydrolysis in Ntn amidotransferases has been subject to investigation. The importance of the terminal cysteine has been established. Nucleophilic attack by the thiol leads to formation of a covalent intermediate. Prior deprotonation of the cysteine by a base is necessary but the identity of this base was unclear. A sequence alignment between GAT domains of PURF, GLMS, asparagine synthetase and glutamate synthase, which have 45-50% identity, lead to the hypothesis of the existence of a catalytic Cys-His-Asp triad (Mei & Zalkin, 1989). In support of this theory, mutation of His 101 in PURF resulted in loss of activity (Mei & Zalkin, 1989). However the crucial His was missing from the sequences of some active asparagine synthetases (Boehlein *et al.*, 1994). Mutational analysis of the GAT domain of GLMS highlighted the necessity of the N-terminal cysteine but found that the mutation of each of the four histidine residues which are conserved in the Ntn family failed to decrease the hydrolysis rate (Badet-Denisot & Badet, 1996). The structure of PURF at 3 Å resolution (Smith *et al.*, 1994), showed neither histidine, nor any other suitable residue is in a position to abstract a proton from the cysteine. It was proposed that the N terminal α-amino group could fulfil this role (Smith, 1995).

Glucosamine 6-phosphate synthase is anomalous in its inability to utilise nascent ammonia in place of amide from glutamine for the amination of fructose-6-phosphate, which suggests that coupling between the domains of the substrate bound enzyme must be exceptionally tight, shielding the amide acceptor site on the synthase domain from the surroundings.

Glutamine analogues, including DON, iodoacetamide and fumaroyl-di-propionic acid derivatives, notably the methyl ester, FMDP, alkylate the N-terminal cysteine resulting in loss of glutamine hydrolysis activity. In the presence of fructose-6-phosphate, addition of 1 equivalent of DON is required for complete loss of activity, while, without fructose 6-phosphate, 0.5 equivalents of DON are sufficient, indicating that there is negative co-operativity between the two domains. These inhibitors have been found to possess antifungal/ antibacterial activity.

The *E. coli* enzyme is a dimer of 67 kDa subunits. Each subunit comprises N-terminal GAT and C-terminal synthase domains, connected through a central hinge region which is uniquely cleaved by chymotrypsin at Tyr 240. A mixture of isolated GAT and synthase domains has no glucosamine-6-phosphate synthesising activity, however GAT domains bind glutamine and retain around 7% of their glutaminase activity. The isolated synthase domain binds 0.16 equivalents of fructose 6-phosphate, compared to 1 equivalent for the intact enzyme.

Figure 1.

Inhibition of glucosamine 6-phosphate synthase by fumaroyl di-propionic acid derivatives

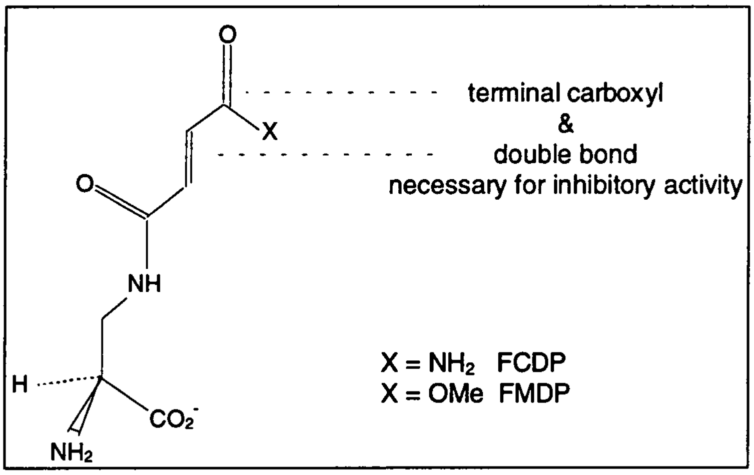
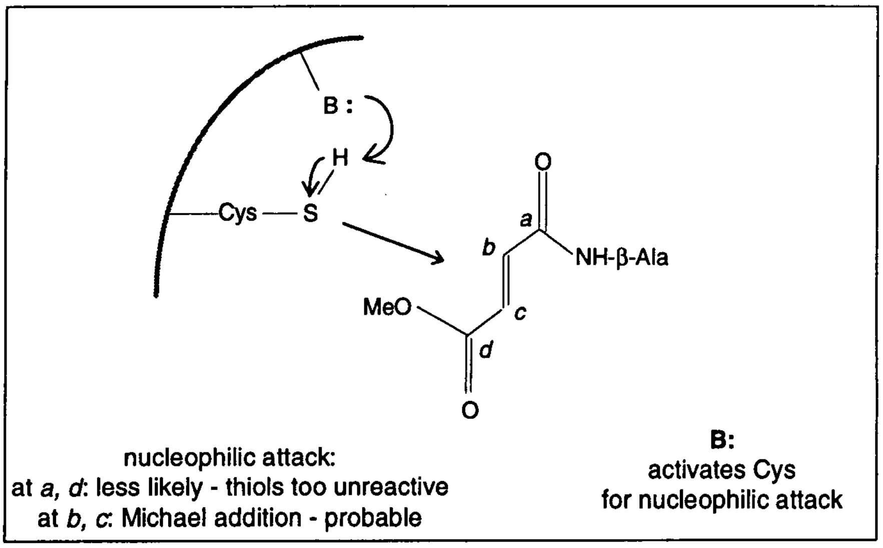


Figure 2. Possible mechanisms proposed for formation of Cys1 - FMDP complex



Overview

Genes encoding the two separate domains of Glucosamine-6-phosphate synthase were cloned and the proteins were overexpressed and purified by Obmolova *et al.* (1994). The determination of crystallisation conditions and initial collection of native data were also accomplished before the start of this PhD project, by Galya Obmolova and Alexei Teplyakov, although these steps are outlined below. The remainder of the work described in the experimental sections which follow was performed as part of the PhD. The appendix on the structure of the GAT domain which concludes the chapter is the work of Michail Usupov and Alexei Teplyakov. It is added for general information, since the results are extremely interesting from both crystallographic and biological points of view and also as an indication of how far the work described in the experimental section was justified and where it was misguided.

Experimental, GAT Domain

Crystallisation and data collection

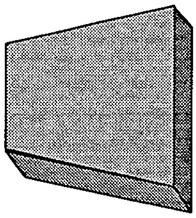
Table 2. Native data collection and processing, GAT domain

parameter	*value*
resolution (Å)	1.8
measured reflections	201 589
In $P2_12_12_1$:	
unique reflections	46 503
R_{merge} (%)	6.1
completeness (%)	98.5
In $P2_1$:	
unique reflections	78 923
R_{merge} (%)	6.2
completeness (%)	86.6

Crystals were obtained using hanging drop vapour diffusion. They were grown at 4 °C from a 1 M sodium acetate solution, with 20% PEG 4000 and 0.1 M cacodylate buffer at pH 6.5. 100 mM of the substrate, glutamine, was present in the crystallisation solution. Crystals grew in 2-4 weeks, reaching a size of 0.5 mm x 0.5 mm x 0.2 mm, Figure 3. The crystals diffracted to 1.8 Å. The point group symmetry was determined as P222, with cell dimensions *a*= 70.4 Å , *b* = 82.5 Å and *c* = 86.1 Å. This gives 2 molecules in the asymmetric unit with 48% solvent content. Systematic absences along all axes indicated the space group $P2_12_12_1$. Data were collected on the X31 EMBL beamline, using a MAR

Research imaging plate detector. The data were processed using DENZO and SCALEPACK (Otwinowski & Minor, 1993). Statistics are listed in Table 2.

Figure 3. GAT domain crystal shape



The native Patterson map contained a strong peak, with height 35% of the origin peak, at (0.50,0.53,0). A self-rotation search found no significant peaks. This indicated that the two molecules in the asymmetric unit had essentially identical orientations and were related by a translation of approximately (0.5,0.5, 0). If this symmetry had been exact this would have been a C-centred lattice. The diffraction pattern contains a complex system of pseudo absences, as described and illustrated in Figure 4. This phenomenon is caused by the divergence from ideal symmetry along the *b* axis.

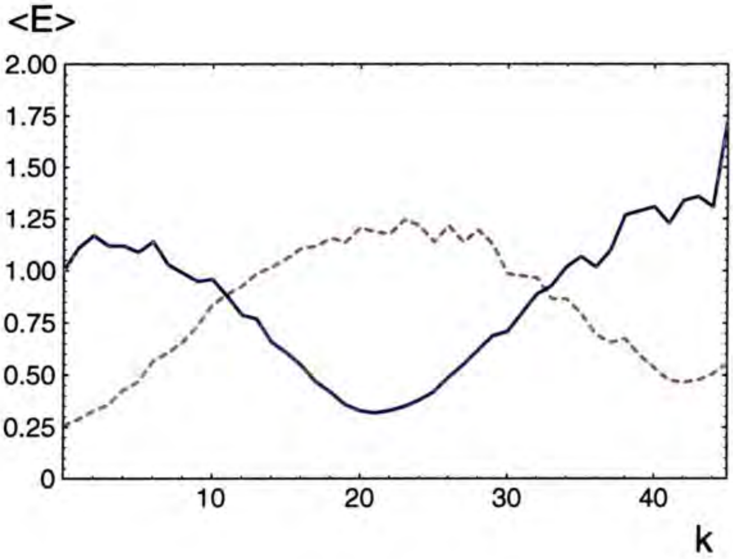


Figure 4.

The sinusoidal pattern of mean intensity for h+k reflections with increasing k displayed by the form I data.

<E> was plotted against k for h+k=2n (blue, continuous line) and h+k=2n+1 (red, dashed line) reflections.

This pattern corresponds to pseudo-absence of h+k=2n+1 reflections for low k(0-4), h+k=2n for medium k(20-24) and h+k=2n+1 again for high k(40-44). The wavelength of this pattern is related to the difference between the pseudo-translation and the translation giving an exact, C-centred cell.

Multiple isomorphous replacement

An extensive search for isomorphous heavy atom derivatives was performed. Data collections on prospective derivative crystals are summarised in Table 3. These crystals were reasonably robust, remaining intact under a variety of soaking conditions and often diffracting well, to 2.7 Å resolution or better. In some cases, merging statistics with native data appeared promising. However, the only significant feature which could be seen on the isomorphous and anomalous Patterson maps was the peak at (0.5, 0.53, 0) resulting from the translation between the two molecules in the asymmetric unit. Thus, it proved impossible to locate any heavy atom sites and it remained unclear whether any of the crystals tested were actually derivatives.

Table 3. The search for derivatives of GAT domain crystals

compound	*concentration (Mm)*	*soak time (days)*	*resolution (Å)*	*R_{merge} /high resolution (%)*	*R_{native} (%)*
mercury					
$HgAc_2$	0.1	1	2.3		18.3
$HgAc_2$	0.5	3	2.25	5.9 / 16.8	15.1
DTT/$HgAc_2$	6 / 0.5	3hrs/1day	2.7	5.1	21.9
DTT/$HgAc_2$	6 / 0.45	3hrs/1day	2.6	9.3 / 20.0	20.8
EMTS	0.5	1	2.25	7.0 / 18.7	10.4
EMTS	2	2	2.15	3.2 / 7.0	9.9
EMTS	2	1	2.7	3.3 / 10.2	21.3
EMTS	1	1	2.4	**monoclinic, $P2_1$**	
transition metal					
K_2PtCl_4	2	2	2.35		27.3
K_2PtCN_4	3	7	2.2		8.5
$Pt(NH_3)_2(NO_3)_2$	3	3	2.3	6.2 / 18.7	13.4
$Pt(NH_3)_4Cl_2$	1	21	2.6	6,3 / 15.8	16.7
$Pt(NH_3)_4Cl_2$	5	7	2.6	7.2 / 13.0	26.4
K_2PtBr_6	1.5	3	2.7	8.8 / 40.0	36.5
K_3IrCl_6	4	3	2.7	5.0 / 13.3	13.9
lanthanide/ actinide					
$Sm(NO_3)_3$	7	7	2.2		9.0
$SmCl_3$	7	3	2.7	6.5	29.6
UO_2Ac	2	3	2.25		10.0
$UO_2(O_2CH_2)_2$	4	3	2.7	8.0 / 26.0	19.6
EuCl	4.8	7	2.6	6.3 / 18.4	40.0
$La(NO_3)_3$	6	3	2.7	6.1	14.5
ThAc	1.6	3	3	61. / 14.2	35.0

Molecular replacement

PURF from *Bacillus subtilis* has a GAT domain which has 27% sequence identity with that of the *E. coli* GLMS GAT domain. Structure solution by molecular replacement was attempted, using the GAT domain from the structure of PURF at 3 Å resolution (Smith *et al.*, 1994), PDB code 1GPH, as a search model. The model included all atoms of the first 224 amino acids with unaltered B factors. The cross-rotation function with an integration radius of 18 Å and the translation function were calculated in the resolution range 10-3 Å using AMORE (Navaza, 1994). A clear rotation solution was found, with an peak height of 18.6 σ, compared to the next, 14.6 σ. However the translation function gave many peaks, with similar peak heights.

At this stage, the choice of space group was reconsidered. It was suggested that the correct space group could be $P22_12_1$. The systematic absences along the *a* axis can be explained by the fact that the translation component of the intermolecular vector is almost exactly half the unit cell width. Molecular replacement, in this new space group yielded a unique translation solution, with correlation 21% and R factor 62.9%, other solutions having correlations not higher than 18.2%. This solution was fixed and the search for the second molecule followed. Two solutions were obtained with equal figures of merit, correlation 42.7%, R 56.1%. Both solutions were acceptable with respect to packing and close contacts within the unit cell. The two solutions were related by the vector (0, 0.053, 0), the translation between the two peaks in the native Patterson map which arises from pseudo centring. Least-squares refinement with PROLSQ (Konnert & Hendrickson, 1980) was carried out using both solutions as starting models and data in the resolution range 10 -1.95 Å. The R factor dropped from 62.4% to 38.7% and R_{free} from 62.3% to 53.1% in both cases, so the solutions were still indistinguishable. The density maps remained uninterpretable and no further progress in refinement was possible.

Alteration of crystal symmetry

The presence of a pseudo-symmetric translation in the unit cell with the consequent weakness of the pseudo-absent data and the large peak in the Patterson map resulted in a loss of information which hindered progress towards structure solution. Attempts were made to grow crystals with a different lattice, lacking the pseudo-symmetric translation.

Alterations were made to the 'normal' crystallisation conditions, which were 0.8 - 1.2 M NaAc, 20% PEG 4000, 0.1 M cacodylate buffer, pH 6.5. Options considered included changing the cation, the anion, the type of PEG or the pH and the addition of organic additives to the crystallisation solution. Crystals were grown under new conditions and their symmetry determined, with the results listed in Table 4.

Table 4. The search for a new symmetry for GAT domain crystals

crystallisation conditions (changes)	*lattice: a,b,c (Å)*	*comments*
PEG 8000	normal	
PEG 8000, 3% MPD	primitive monoclinic: 53, 87, 56, β = 99°	disordered
PEG 8000, 3% MPD	primitive monoclinic: 87, 71, 168, β = 92° or centred monoclinic: 255, 63, 98, β = 95°	disordered
PEG 8000, 3% ethanol	primitive monoclinic: 53, 87, 56, β = 101° or centred orthorhombic: 69, 83, 87	disordered
PEG 8000, 3% ethanol	primitive monoclinic: 53, 87, 56, β = 99° or centred orthorhombic: 71, 83, 87	disordered
PEG 8000, 1% cyclo-hexane	normal	
PEG 6000	normal or primitive orthorhomic: 83, 87, 42	disordered
PEG 6000 β-octyl glyceride	normal	

The presence of ethanol and MPD during crystallisation evidently affected the lattice, but the crystals produced were disordered. Alteration of the crystal lattice by soaking crystals grown under normal conditions in concentrated salt solution (3.9 M NaAc), 5% ethanol, 55% PEG 4000 and 60% MPD was also tried. Crystals were found to diffract following this treatment, although to a lower resolution limit, but no lattice changes were observed.

Synthase Domain

Crystallisation and data collection

Crystals were grown by hanging-drop vapour diffusion, at 4 °C, from a solution containing 27-31% sodium formate buffered by 0.1 M imidazole at pH7. 5 mM fructose-6-phosphate, the substrate, was present in the solution. Crystal growth in the absence of substrate is also possible. Hexagonal bipyramidal crystals grew in 2-8 weeks with dimensions 0.6 mm x 0.3 mm x 0.3 mm. These crystals diffracted to a limit of 2.2 Å. The data indexed in space group $P6_1$, or its enantiomer, $P6_5$, with $a = b = 63.5$ Å and $c = 334.5$ Å. Two molecules in the asymmetric unit give a solvent content of 49%. Data were collected on the X11 EMBL beamline to 2.6 Å resolution and processed using DENZO and SCALEPACK. 108 620 measured reflections gave 23 992 unique data, with R_{merge} 5.4%. The native data are summarised in Table 5.

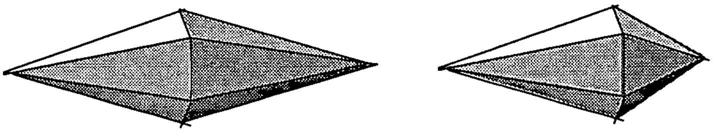
Structure solution attempts

Structure solution by multiple isomorphous replacement was attempted and data were collected on potential derivative crystals as listed in Table 6. R_{merge} was high for all derivative datasets and the scaling of native and prospective derivative data proved impossible. The merging statistics for each pair of datasets are listed in Table 7.

The self rotation function for each dataset was calculated, Table 8. The equatorial angle, ϕ, which is equivalent to the angle of rotation about the 6_1 axis between the two molecules in the asymmetric unit, appears to vary widely between datasets. For a single dataset, Pt2 as defined in Table 6, the images were reprocessed in three sets of 20°, chronologically grouped. R_{merge} for these subsets was lower than for the whole dataset and the self rotation function gave a different value for ϕ in each case, Table 9. From this variation in the rotation function, it appears that the alignment of the lattice varies between crystals and either with orientation or time or both, for a single crystal during data collection. The most obvious interpretation is that there is some type of rotational disorder in the lattice.

If $\phi = 0$, the space group would be $P6_122$. The symmetry of these crystals may therefore be considered to be disordered $P6_122$. Progress in structure determination may be possible if this disorder could be removed, either by an induced lattice transformation, for instance by soaking or freezing of the crystals, or by a change in crystallisation conditions. The crystals are hexagonal bipyramidal, but some are symmetrical about a central mirror plane, while others are not, Figure 5. It was suggested that crystals with this symmetry may possess a $P6_122$ lattice, while the asymmetric ones were in $P6_1$.

Figure 5. Synthase domain crystal shapes



symmetric - $P6_122$?   asymmetric - $P6_1$?

Alteration of crystal symmetry

New crystallisation conditions, resulting in crystals of altered symmetry, were sought. Crystals were found to grow in a range of conditions, including: 28% sodium formate precipitant, buffered by imidazole at a range of pHs, 7.0 - 8.0, and concentrations 0.1 - 0.15 M, or 0.1 M HEPES at pH 7.0 - 8.0; 1 M sodium acetate, with 0.1 M TRIS buffer, at pH 8.5 and 10% PEG 4000, 35-40% ammonium sulphate, at pH 7.0; 38% ammonium sulphate with 0.1 M Imidazole at pH7.0; 35% sodium citrate with 0.1 M HEPES at pH7.0; 35-40% lithium sulphate with 0.1 M HEPES at pH 7.5. Generally, the crystals appeared to have the same form, and those tested possessed the old unit cell. The use of sodium citrate as precipitant resulted in hexagonal prisms. However the diffraction from these was too poor for the unit cell to be ascertained.

Table 5. Native datasets collected, synthase domain

dataset	resolution (Å)	$P6_1$ a,c (Å)		N_{unique}
GS7	10 - 2.5	63.7	335.4	15 373
GS8	18 - 2.6	63.6	335.0	22 619
GS9_*h*	15 - 2.6	63.5	334.3	23 992
GS9_*l*	29 - 3.3	64.0	336.5	11 960
GS10	25 - 2.6	64.2	337.8	22 644
GS11	35 - 5.6	63.7	335.6	2 279

Table 6. 'Derivative' datasets collected, synthase domain

dataset	heavy atom compound	conc (mM)	soak time	resolution (Å)	symmetry, cell $P6_1$ a,	 c (Å)	R_{merge} (high res) (%)	complete (high res) (%)	N_{unique}
Pt1	$PtPy_4Cl_2$ $.2H_2O$	1.15	4 hrs	20 - 2.6	64.7	340.2	12.7 (42)	99.3 (99.2)	24 431
Pt2	K_2PtCl_4	1	1 day	20 - 2.6	64.2	340.7	10.0 (35)	98.4 (100)	24 084
Pt3	$PtPy_4Cl_2$ $.2H_2O$	1	4 hrs	35 - 5.6	63.6	334.1	6.6 (10.5)	89.0 (84.1)	2 089
Hg1	ETMS	0.39	1 hr	20 - 2.6	cracked		-	-	-

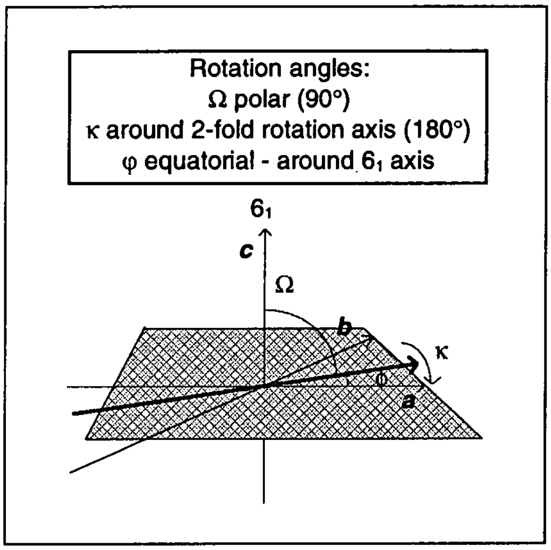
Table 7. R_{merge} for pairs of synthase datasets

	GS7	GS8	GS9_*l*	GS9_*h*	GS10	GS11	Pt1	Pt2	Pt3
GS11	30.0	23.6	29.5	34.6	30.4	-	25.5	26.2	52.7
GS11a*	54.0	40.6	50.1	41.6	42.8	-	30.0	30.0	32.3
Pt1	40.0	29.6	41.1	34.7	36.4		-	11.9	37.8
Pt1a*	42.2	31.1	42.1	35.0	37.5		-	12.3	36.1
Pt2	41.2	28.9	40.9	33.7	36.8			-	37.9
Pt2a*	42.8	30.6	42.5	34.3	37.9			-	35.9
Pt3	49.3	37.7	46.8	30.9	36.9				-
Pt3a*	17.6	20.0	42.1	20.2	13.5				-

* Gsa ≡ enantiomer of GS

Table 8. Peaks in the self-rotation function for synthase data sets

dataset		peak height (% origin)	ϕ (°)*	Ω (°)	κ (°)
GS7	1	83.8	-6.2	90	180
	2	33.9	0.0	90	180
GS8	1	81.4	-14.1	90	180
	2	71.6	-3.0	90	180
GS9_l	1	71.7	-5.3	90	180
	2	69.2	0.0	90	180
GS9_h	1	96.0	-0.7	90	180
GS10	1	76.1	12.5	90	180
GS11	1	84.8	-9.1	90	180
	2	79.0	0.0	90	180
Pt1	1	97.2	-0.7	90	180
Pt2	1	95.7	+0.5	90	180
Pt 3	1	89.0	+5.9	90	180
	2	80.3	0.0	90	180


Rotation angles:
Ω polar (90°)
κ around 2-fold rotation axis (180°)
φ equatorial - around 6₁ axis
6₁
c
Ω
b
κ
φ
a

* $P6_1$ is a polar space group, so re-indexing with the 6_1 axis in the opposite direction will cause ϕ to change sign. The rotation function was calculated, as well as could be determined, for all data indexed the same way round, i.e. for the sets of data which gave the best merging statistics.

Table 9. Results of processing dataset Pt2 in three subsets

images	$\phi_{\text{spindle axis}}$ range (°)	R_{merge} (%) (high resolution.)	rotation function peak height (% origin)	$\phi_{\text{rotation function}}$ (°)
1-59	60	10.0 (35)	95.7	+0.5
1-13	20	6.8 (18.6)	92.6	+5.2
13-45	20	7.0 (28)	87.5	-4.5
47-59	20	5.8 (27)	83.4	-10.9

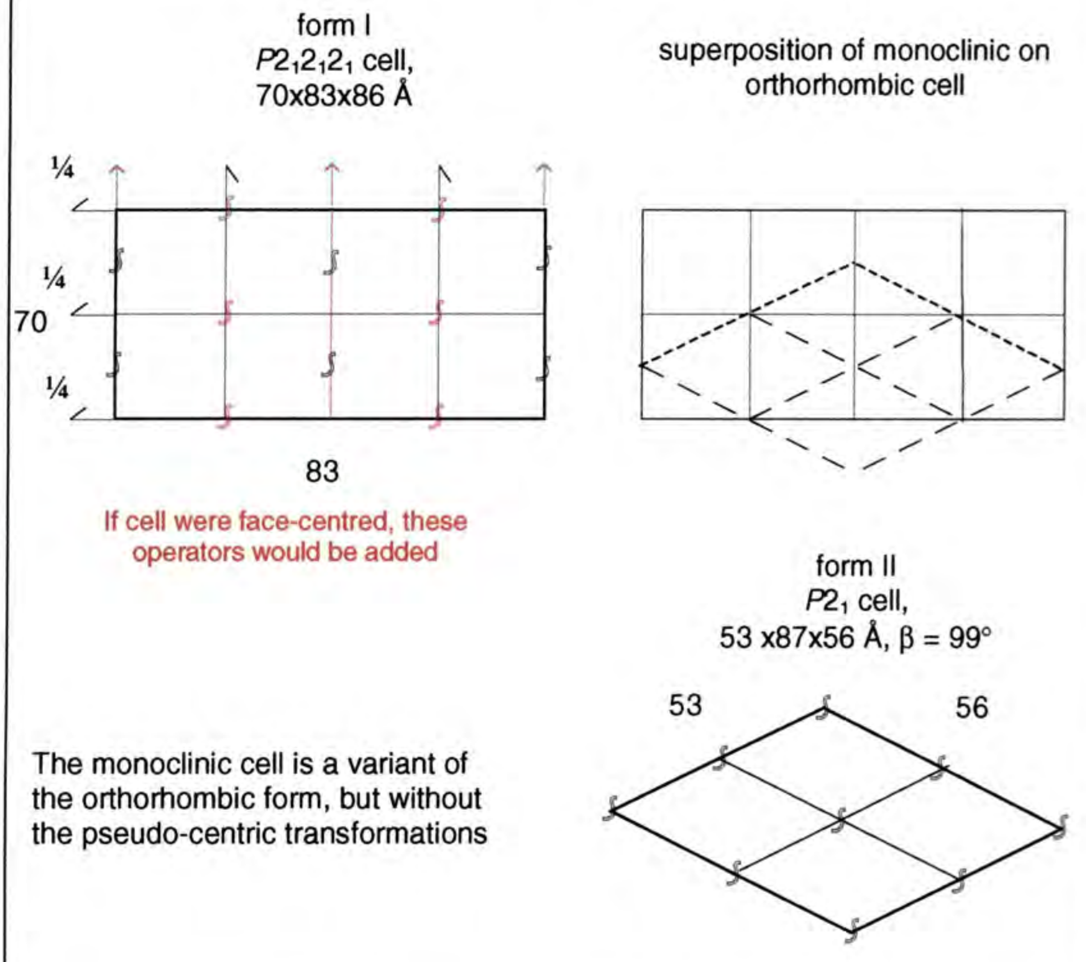
Appendix: Structure of the GAT domain

(Summary of work by Usupov & Teplyakov)

Solution and refinement

During the screening for derivatives, data were collected on a single crystal which indexed in the monoclinic space group, $P2_1$, with a = 53.5 Å, b = 87.3 Å, c = 56.6 Å, β = 98.9 °. In the monoclinic cell, form II, two molecules are related by a non-crystallographic 2-fold axis. This new form has a similar packing motif to the orthorhombic cell, form I, but without the pseudo symmetric operators, Figure 6.

Figure 6. Relationship between GAT domain crystal forms I and II



If cell were face-centred, these operators would be added

The monoclinic cell is a variant of the orthorhombic form, but without the pseudo-centric transformations

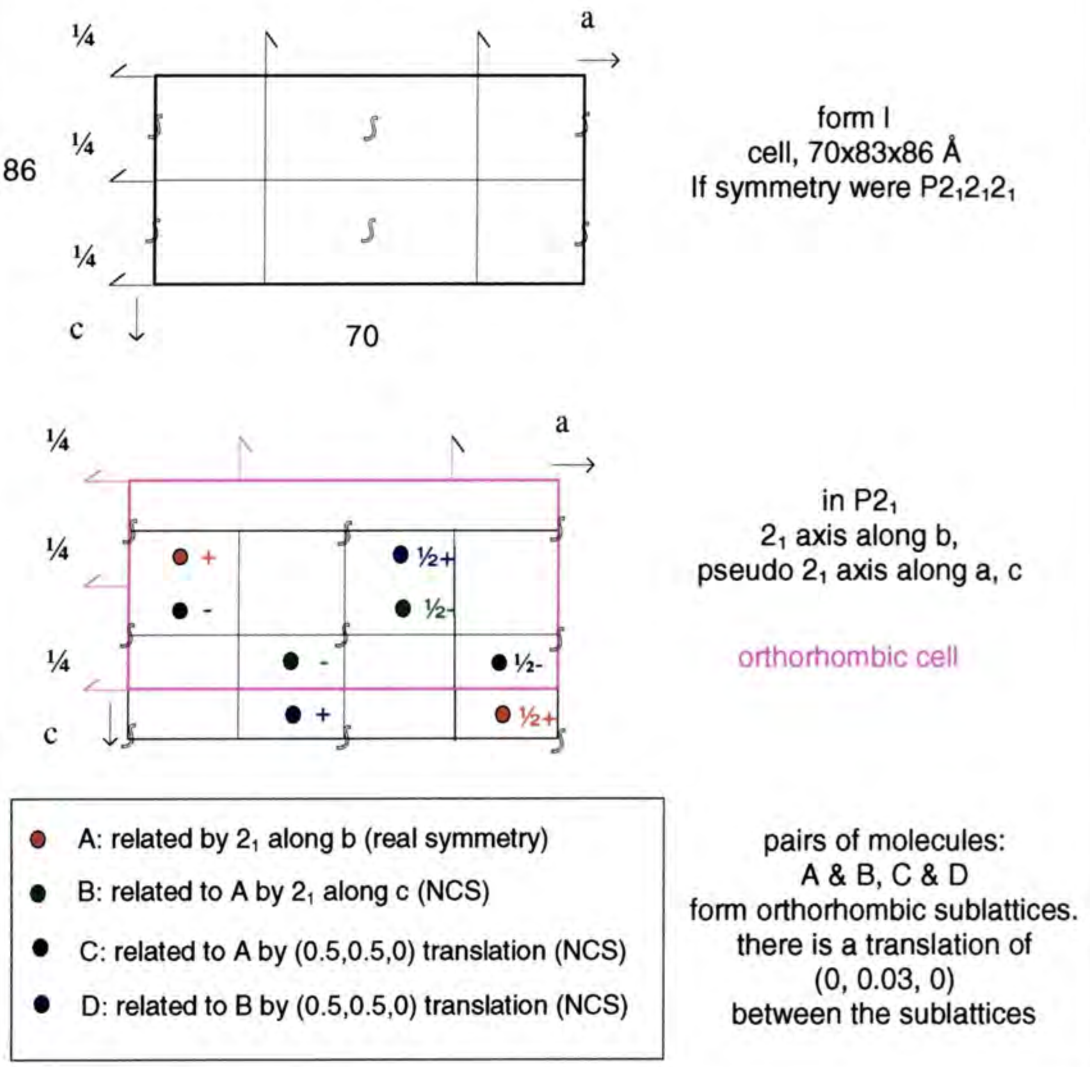
Structure solution by molecular replacement was tried unsuccessfully, using the form II data with the PURF GAT domain as a search model, as described above for form I. An attempt was then made using the preliminarily refined molecular replacement solution in $P22_12_1$ as a search model. This yielded a solution for two molecules with correlation 57.8% and R 46.5%. The solution was subjected to refinement at 23 - 2.4 Å resolution using a maximum likelihood approach, REFMAC (Murshudov *et al.*, 1996), with the application of strict NCS restraints. 3% of the data were separated from the working data and used to evaluate R_{free}. Phase improvement, using the CCP4 program DM, resulted in a density map in which the deviations of the chain from the search model could be traced, allowing manual corrections to be applied, using O (Jones *et al.*, 1991). This model refined to R 20.2% and R_{free} 27.1%.

The extension of the model to higher resolution and the investigation of complexes necessitated a solution of crystal form I. The data collection of form II was serendipitous and numerous crystallisation attempts had failed to procure another such crystal. Molecular replacement in $P22_12_1$ using the refined $P2_1$ model as a search model resulted in two solutions which could not be refined, as had occurred previously. A solution was then searched for in all other primitive orthorhombic space groups, to no avail. The possibility of a lower symmetry was then considered.

A unique solution was finally obtained in $P2_1$, with the 2_1 axis along the 82 Å cell edge. The solution had correlation 83% and R 34.3%, four crystallographically independent molecules; A, B, C, & D, with two pairs of molecules, A&C, B&D, related by the former pseudo-centring vector, and two pairs; A&B, C&D, related by the former orthorhombic 2-fold along the *c* axis, Figure 7. The crystal lattice can be described as a superposition of two orthorhombic sublattices, shifted by 0.027*b* along a common monoclinic 2_1 axis.

Data were reprocessed in $P2_1$, with statistics listed in Table 2. Due to the misapprehension that the symmetry was orthorhombic, only 90° of data had been collected and the completeness in the new monoclinic symmetry, was only 86%. PROLSQ refinement was performed at 10-1.8 Å resolution without NCS restraints. 1% of the data were used for R_{free} calculation. Water sites were added where peaks of height greater than 3.5 σ in the difference map came within bonding distance, 3.5 Å, of potential H-bonding partners. Their occupancies were set to 100% and not refined.

Figure 7. Relationship between monoclinic and orthorhombic symmetries of form I, and between the 4 independent molecules in the monoclinic cell

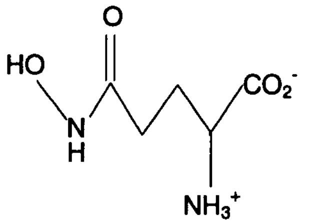

¼
a
¼
86
¼
c
70
form I
cell, 70x83x86 Å
If symmetry were $P2_12_12_1$
¼
a
¼
+
½+
-
½-
¼
-
½-
c
+
½+
in $P2_1$
2_1 axis along b,
pseudo 2_1 axis along a, c
orthorhombic cell
● A: related by 2_1 along b (real symmetry)
● B: related to A by 2_1 along c (NCS)
● C: related to A by (0.5,0.5,0) translation (NCS)
● D: related to B by (0.5,0.5,0) translation (NCS)
pairs of molecules:
A & B, C & D
form orthorhombic sublattices.
there is a translation of
(0, 0.03, 0)
between the sublattices

The final model had R 17.9%, R_{free} 25.4% and 770 solvent sites, 238 residues modelled for A & C and 239 for B & D, out of 240 expected. A metal ion site was detected in the density. It is co-ordinated by five oxy-ligands, with distances in the range 2.2-2.6 Å. The metal was modelled as Na^+, since this cation was present in high concentration in the crystallisation medium. The substrate glutamine was also present in the solution and difference density for this molecule was observed and modelled. Substrate binding is discussed below.

An inhibitor complex

Data were collected on a crystal which had been soaked in a 50 mM solution of L-Glu-hydroxamate, Figure 8, a competitive inhibitor of Gln binding and hydrolysis. The inhibitor binding was identified in the difference map and the structure refined, using the model of the substrate complex.

Figure 8. L-Glu hydroxamate, an inhibitor of Gln hydrolysis



Description of the structure

The four layer structure consists of a pair of β-sheets sandwiched between α-helical zones, Figure 9. The central strands of the β-sheet assembly are planar, but the end strands twist, so that opposite ends of the sheet lie perpendicular to one another. The C-terminal strands, β14 and β15, can be considered as a single strand forming the closing link between the β-sheets. The secondary structural units are joined by short loop regions, 13 out of 18 of which contain Gly residues in conformations forbidden to other residues. These Gly residues are highly conserved in the Ntn family of amidotransferases. The region between the sheets is packed with hydrophobic residues. The majority of α-α and α-β interactions are also hydrophobic. The Na^+ binding site is formed from loops connecting α5 to β7 and β8 to β9.

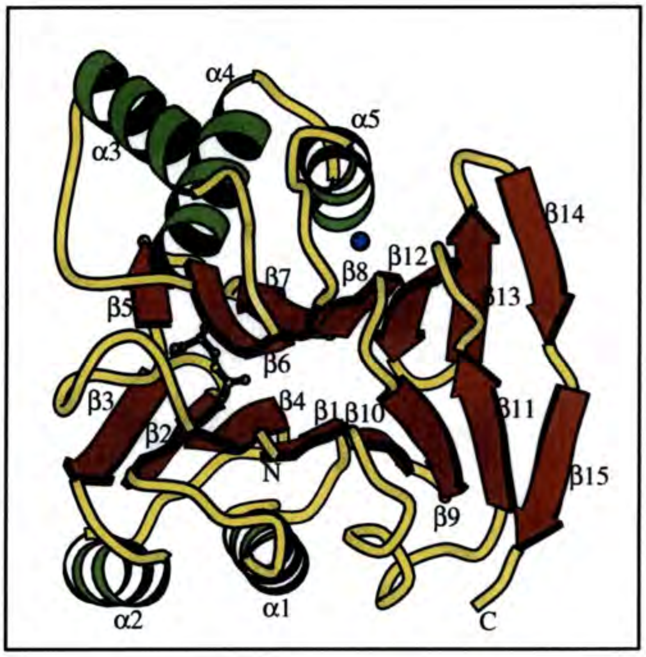


Figure 9.

Ribbon representation of the GAT domain of GLMS produced using MOLSCRIPT (Kraulis, 1991). Alpha-helices are depicted in green, beta-strands in red. A ball and stick representation of the product, Glu, lies in the active site.

Substrate binding site

The catalytic residue, Cys1, is situated at the bottom of a deep pocket in the centre of the sheet structure. The substrate lies within this pocket. The binding pocket is comprised principally of the loop sections connecting β4 to β5 and β6 to α3. Numerous H bonds are formed between ligand and protein. The binding pocket is enclosed by the loop residues 73-79, which interact with the α-CO_2^- and α-NH_3^+ of the molecule, the other end of which points towards the catalytic site at the bottom of the pocket. H-bonds are formed between the γ-carboxyl group and Cys1 N, Trp74 N and Gly99 N, which are all proton donors. This provides evidence that the ligand bound is the product, Glu, rather than Gln. In addition, the refinement of both γ-carboxyl atoms as oxygens gives equal B values for the pair, while, if one is modelled as nitrogen, the nitrogen B factor drops by 5 $Å^2$. It is predicted that the amide group of Gln would lie less deep within the pocket, forming an H-bond to Trp74 O. The amide of the inhibitor L-Glu hydroxamate is indeed observed to interact in this manner.

Comparison to PURF structure, the loop lid on the binding site

The structures of the GAT domains of PURF (Smith *et al*, 1994) and GLMS are very similar, the main structural difference being that the former lacks the C-terminal pair of β-sheets. However, superposition of the 219 common CA atoms gives an rmsd of 2.15 Å, due to the considerable deviation of almost all the loop regions between the two structures. Superposition of 70 CA atoms from β-sheet regions gives an rmsd of only 0.93 Å.

The maximum deviation of up to 10 Å occurs for the CA atoms of residues 74-83, the loop region joining β4 and β5, which forms a lid over the Gln binding site. In the PURF structure, the binding site is empty and the lid loop in open conformation, with rather high B factors indicating considerable motion. In the GLMS structure, the occupied site is closed by the lid, which is fixed in a much more rigid conformation. The open lid appears to provide the only access to the binding site. A gate mechanism, mediated by interaction with the synthase domain, is proposed for lid opening. In the PURF structure, the Gln binding site is shielded from the bulk solvent by the synthase domain and relative movement of the two domains is necessary for the opening of the site. In the structure of PURF complexed with the glutaminase inhibitor DON (Kim *et al.*, 1996), the loop also adopts the closed conformation.

Mechanism for glutamine hydrolysis

From the structure of the enzyme bound to the reaction product, a mechanism for glutamine hydrolysis can be deduced, Figure 10.

1/ Deprotonation of the Cys1 thiol is accomplished with the N-terminal α-amino group acting as a base catalyst, via a bridging water molecule.

2/ Nucleophilic attack of the activated thiol on the amide carbon of Gln forms a tetrahedral intermediate, stabilised by H-bonding with Asn 98 ND and Gly 99 N.

3/ The intermediate collapses, releasing ammonia and forming a γ-glutamylthioester.

4/ Deacylation results from the attack of a water molecule, activated by the N-terminal α-amino group, in the same manner as in step 1, to form a second tetrahedral intermediate, stabilised as in step 2, followed by the collapse of this intermediate, releasing glutamate.

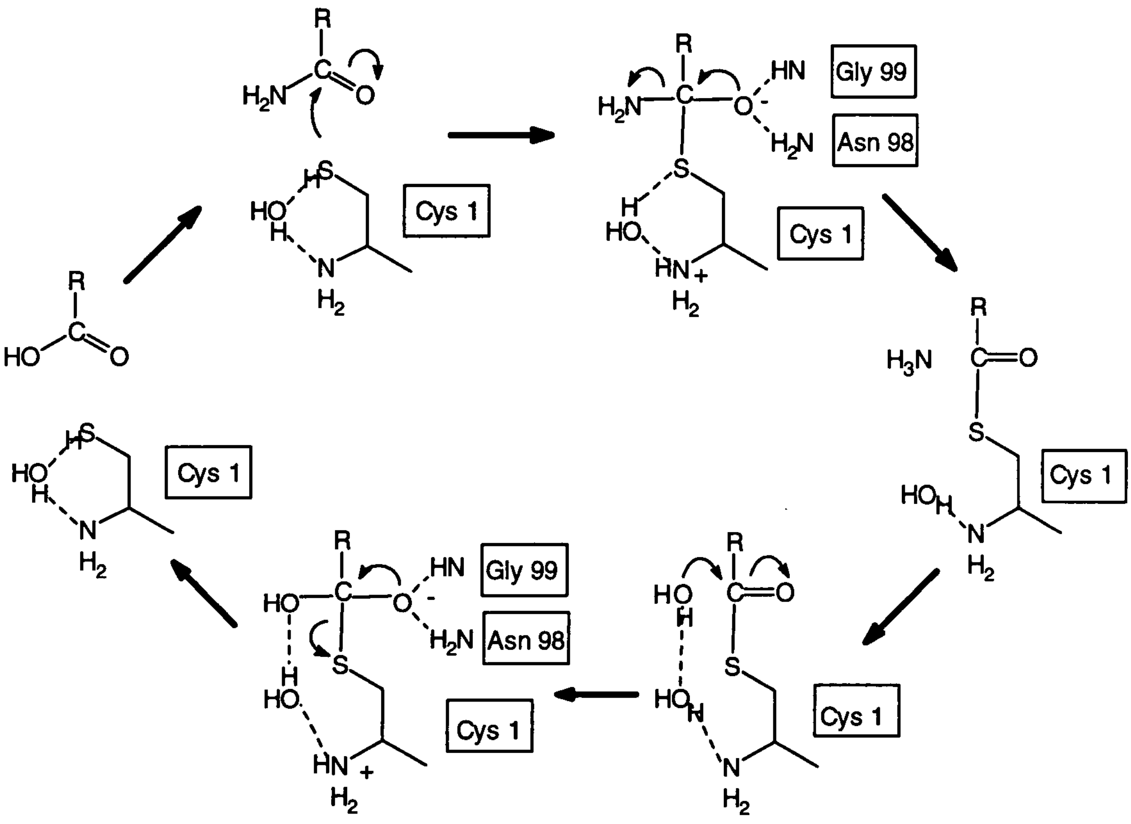


Figure 10.
Proposed catalytic mechanism of Gln hydrolysis by amidotransferases.
Residue numbers are those of GLMS.

The thiol is activated for nucleophilic attack by the N-terminal α amino group. There are no other suitable basic groups neighbouring the active site, in either PURF or GLMS structures. There is no evidence for the involvement of a catalytic Cys His Asp triad of residues, as previously postulated. The evidence suggests that a mechanism of glutamine hydrolysis entailing single amino acid catalysis could be a common feature for the Ntn amidotransferase family of enzymes.

The mechanism assumes the prior repositioning of two key residues. The sidechain of Asn 98, which is hydrogen-bonded to Cys1 N in the GLMS structure should rotate 100° to form, with Gly 99, the oxyanion hole which stabilises tetrahedral reaction intermediates. The thiol of Cys1 points outside the binding pocket in the GLMS structure. A rotation of 180° around the CA - C bond and 120° around CA - CB moves it into the active position, ready for nucleophilic attack. This is the conformation adopted in the PURF structure. This switch is impossible with the lid loop in closed position, therefore, glutamine hydrolysis is dependent on Cys1 adopting the active conformation prior to Gln binding and closure of the pocket.

It has been shown that, for the Ntn family, Gln hydrolysis is inhibited by the reaction product, Glu, in both the isolated GAT domain and the complete enzyme in absence of the N acceptor substrate. This suggests that binding of the N acceptor to the synthase domain results in the co-operative activation of the GAT domain, with rotation of Cys1 into the attacking position.

The unusual inhibition of the GLMS GAT domain by the product, Glu, can now be rationalised. In the isolated domain, Cys1 is free to adopt either active or inactive conformation, at random. Gln in the crystallisation solution is converted into Glu by enzyme molecules in active conformation. Due to the unfavourable interaction with the thiol in the active conformation, Glu complexes preferentially with enzyme in the inactive conformation, trapping molecules in this alignment, since there is no release mechanism.

The existance of a family of Ntn hydrolases has recently been recognised, members of which include penicillin acylase(PA), the proteasome(PROT) and aspartyl glycosylaminase(AGA) (Brannigan *et al.*, 1995). These enzymes are characterised by a common mechanism for hydrolysis, the intial step of which is the nucleophilic attack on the substrate carbonyl carbon by the side chain of the N-terminal residue, Cys for GAT, Ser for PA, Thr for PROT and AGA. This N-terminal side chain is activated by its own α-amino group, via a bridging water molecule.

The Ntn fold is a 4-layer α+β structure incorporating two antiparallel β-sheets, one flat the other twisted. Residues from the twisted sheet are incorporated in cavities in the flat sheet, thus the two sheets are packed very closely. The structural alignment of active sites of different Ntn hydrolases shows that elements of the catalytic centres are equivalent. Ser B1 OG, Ser B1 N, Ala B69 N and Asn B241 ND2 of PA may be superemposed on Cys1 SG, Cys1 N, Gly103 N and Asn102 N of the PRPP GAT domain, with an rmsd 0.7 Å. This common framework suggests that a common mechanism exists, involving intermediates with the same chirality.

Glutaminase / Ntn References

Badet, B., Vermoote, P., Haumont, P.-Y., Lederer, F. & Le Goffic, F. (1987) Glucosamine synthase from *Escherichia coli*: purification, properties and glutamne-utilising site location. *Biochemistry* **26**, 1940-1948.

Badet-Denisot, M.-A. & Badet, B. (1996) Mutational analysis of the glutamine amide transfer domain of glucosamine 6-phosphate synthase. *In press*.

Boehlein, S.K., Richards, N.G.J. & Schuster, M.S. (1994) Glutamine-dependent nitrogen transfer in *Escherichia coli* asparagine synthetase B. Searching for the catalytic triad. *J. Biol. Chem.* **269**, 7450-7457.

Brannigan, J.A., Dodson, G., Duggleby, H.J., Moody, P.C.E., Smith, J.L., Tomchick, D.A. & Murzin, A.G. (1995) A protein catalytic framework with an N-terminal nucleophile is capable of self-activation. *Nature* **378**, 416-419.

Denisot, M.-A., Le Goffic, F. & Badet, B. (1991) Glucosamine 6-phosphate synthase from *Escherichia coli* yields two proteins upon limited proteolysis: idemtification of the glutamine aminohydrolase and 2R ketose/aldose isomerase-bearing domains based on their biochemical properties. *Arch. Biochem. Biophys.* **288**, 225-230.

Isupov, M.N., Obmolova, G., Butterworth, S., Badet-Denisot, M.-A., Badet, B., Polikarpov, I., Littlechild, J.A. & Teplyakov, A. (1996) Structure of the glutaminase domain of glucosamine 6-phosphate synthase: implications for the mechanism of nitrogen transfer in Ntn amidotransferases. *Structure.* **4**, 801-810.

Kim, J.H., Krahn, J.M., Tomchick, D.R., Smith, J.L. & Zalkin, H. (1996) Structure and function of the gluamine phosphoribosylpyrophosphate amidotransferase glutamine site and communication with the phosphoribosylpyrophosphate site. *J. Biol. Chem, in press.*

Kucharczk, N., Denisot, M-A., Le Goffic, F. & Badet, B. (1990) Glucosamine 6-phosphate synthase from *Escherichia coli*: determination of the mechanism of inactivation by N^3-fumaroyl-L-2,3-diaminopropionic derivatives. *Biochemistry* **29**, 3668-3676.

Mei, B. & Zalkin, H. (1989) A cysteine-histidine-aspartate catalytic triad is involved in the glutamine amide transfer function in *purF*-type glutamine amidotransferases. *J. Biol. Chem.* **264**, 16613-16619.

Obmolova, G., Badet-Denisot, M-A., Badet, B. & Teplyakov, A. (1994) Crystallisation and preliminary X-ray analysis of the two domains of glucosamine 6-phosphate synthase. *J. Mol. Biol.* **242**, 703-705.

Tesner, J.J.G., Klem, T.J., Deras, M.L., Davisson, V.J. & Smith, L.S. (1996) The crystal structure of GMP synthetase reveals a novel catalytic triad and is a structural paradigm for two enzyme families. *Nature Structural Biology* **3**, 74-86.

Smith, J.L., Zaluzec, E.J., Wery, J.-P., Niu, L., Switzer, R.L., Zalkin, H. & Satow, Y. (1994) Structure of the allosteric regulatory enzyme of purine biosynthesis. *Science* **264**, 1427-1433.

Smith, J. L. (1995) Structures of glutamine amidotransferases from the purine biosynthetic pathway. *Biochem. Soc. Trans.* **23**, 894-898.

Zalkin, H. (1993) The amidotransferases. *Adv. Enzymol. Relat. Areas Mol. Biol.* **66**, 203 - 309.

see also general crystallographic references, end of Chapter 5

Chapter 4:

Anisotropic refinement of two small protein structures

Introduction: Least-squares refinement

The relationship of the parameters of an atomic model to the crystallographic diffraction pattern is well understood. This allows a model to be refined to match experimental diffraction data with high fidelity, using the method of least-squares minimisation.

Given a set of independent observations, $y_1,...y_m$, from which a set of parameters, $x_1,...x_n$. are to be determined, if the observations are linearly related to the parameters, equations (1) can be written.

$$a_{11}x_1 + ...+ a_{1n}x_n = y_1$$

$$a_{m1}x_1 + ...+ a_{mn}x_n = y_m \qquad (1)$$

If m=n, there is an exact solution, but no estimate of the errors on the parameters. If m > n, the equations are overdetermined and a best fit solution can be found by minimisation of the residual, d_i for each observation, where

$$d_i = \mathrm{I}\ y_i - a_{i1}x_1 + ...+ a_{in}x_n\ \mathrm{I} \qquad (2)$$

This can be achieved by the minimisation of the sum of the squares of the residuals, (3) but only if all the observations are equally reliable, otherwise the residuals must be weighted, (4).

$$M = \Sigma_{i=1}^{m} d_i^2 \qquad (3)$$

$$M = \Sigma_{i=1}^{m} w_i\, d_i^2 \qquad w_i = 1/\sigma^2, \qquad \sigma = \text{standard nucertainty of } y_i \qquad (4)$$

M is minimum when the differential with respect to x_j is zero (5). This gives the normal equation for each x_j, (6).

$$1/2\ \partial M/\partial x_j = \Sigma_{i=1}^{m} w_i\, a_{ij}\ (\, a_{i1}x_1 + \ldots + a_{in}x_n - y_i) = 0, \qquad j=1 \text{ to } n \tag{5}$$

$$\Sigma_{i=1}^{m} w_i a_{i1} a_{ij} x_1 + \ldots + .\Sigma_{i=1}^{m} w_i a_{ij}^2 x_j + \ldots + \Sigma_{i=1}^{m} w_i a_{in} a_{ij} x_n = \Sigma_{i=1}^{m} w_i\, a_{ij} y_i \tag{6}$$

Least-squares refinement is relatively insensitive to gross errors and computationally economical. If the residuals obey a normal distribution and the ratio of observations to parameters is large, the reliability of error estimation is good.

A crystallographic least-squares refinement involves minimisation of structure factor or intensity residuals (7). In this case, the observations are not linearly dependent on the parameters that are to be fitted. As a result, iterative refinement steps are required. The solution of the normal equations ignores higher than first order derivatives. Consequently, errors in the trial function must be small, otherwise the refinement will oscillate or diverge rapidly rather than converging.

$$M = \Sigma_h\ 1/\sigma_h^2\ (|F_h^{obs}| - |F_h^{calc}|)^2 \tag{7}$$

The solution of the normal equation is achieved by expressing the problem in terms of matrices. A set of m observations, **Y** can be expressed as a function of a set of n parameters, **x**, with m > n.

$$F(\mathbf{x}) = \mathbf{Y} \tag{8}$$

A trial model is found, $\mathbf{Y_c}$. To get better estimates of $\mathbf{Y_c}$, Taylor expansion gives:

$$\mathbf{Y_c} + (\delta F/\delta \mathbf{x})\ \delta\mathbf{x} = \mathbf{Y_o} \tag{9}$$

which can be rearranged:

$$\mathbf{A}\ \delta\mathbf{x} = \Delta\mathbf{Y} \tag{10}$$

$$\mathbf{A^T A}\ \delta\mathbf{x} = \mathbf{A^T}\Delta\mathbf{Y}$$

$$\mathbf{N}\,\delta\mathbf{x} = \mathbf{A}^{\mathbf{T}}\Delta\mathbf{Y} = \mathbf{r}$$

$$\delta\mathbf{x} = \mathbf{N}^{-1}\,\mathbf{r} \qquad (11)$$

where **A**(m,n) is a matrix, $A_{ij} = \delta F_i/dx_{ji}$, **r** is the residual vector $r_i = \Sigma\ (\delta Y_m/\delta x_i)\ .\ (Y_o - Y_c)m$, **N**(n,n) is the normal matrix $\mathbf{A}^{\mathbf{T}}\mathbf{A}$ and δ**x** is the set of parameter shifts. The chief computational demands are for sufficient memory for storage of the ½n(n+1) terms of the symmetric normal matrix and time for ½nm(n+1) operations on this matrix.

Data collected on small molecule crystals to atomic resolution (~1 Å) give a large data to parameter ratio. There are a comparatively small number of atoms in the structure. Direct methods commonly provide an excellent starting model with co-ordinate errors not greater than 0.1 Å. The relatively small number of parameters also means that ½n(n+1) is not a prohibitively large number for computational memory. Thus, full least-squares refinement of atomic co-ordinates and anisotropic thermal parameters is the routine procedure.

As the number of model parameters, n, increases, the demand for computer time and memory for least-squares refinement increases by more than n^2. Thus, for macromolecules the supply of sufficient computational facilities is non-trivial. However, even if advances in computer technology were to render this factor insignificant some fundamental differences would remain.

In a typical protein crystal, loosely packed molecules are interspersed with channels of solvent. The solvent regions contain diffuse density, the overall B factor of the structure is large and, especially at the protein solvent interface, atoms exhibit a high degree of thermal motion and disorder. Consequently, the high resolution data are weak and limited, leading to a low data to parameter ratio, with the refinement tending to become under-determined. A further difficulty relates to the character of the starting models. Structure solution techniques for macromolecules often yield trial models with large errors, of around 0.5 Å in co-ordinates, and the nature of the errors inherent in a solution differs with the method by which it was derived.

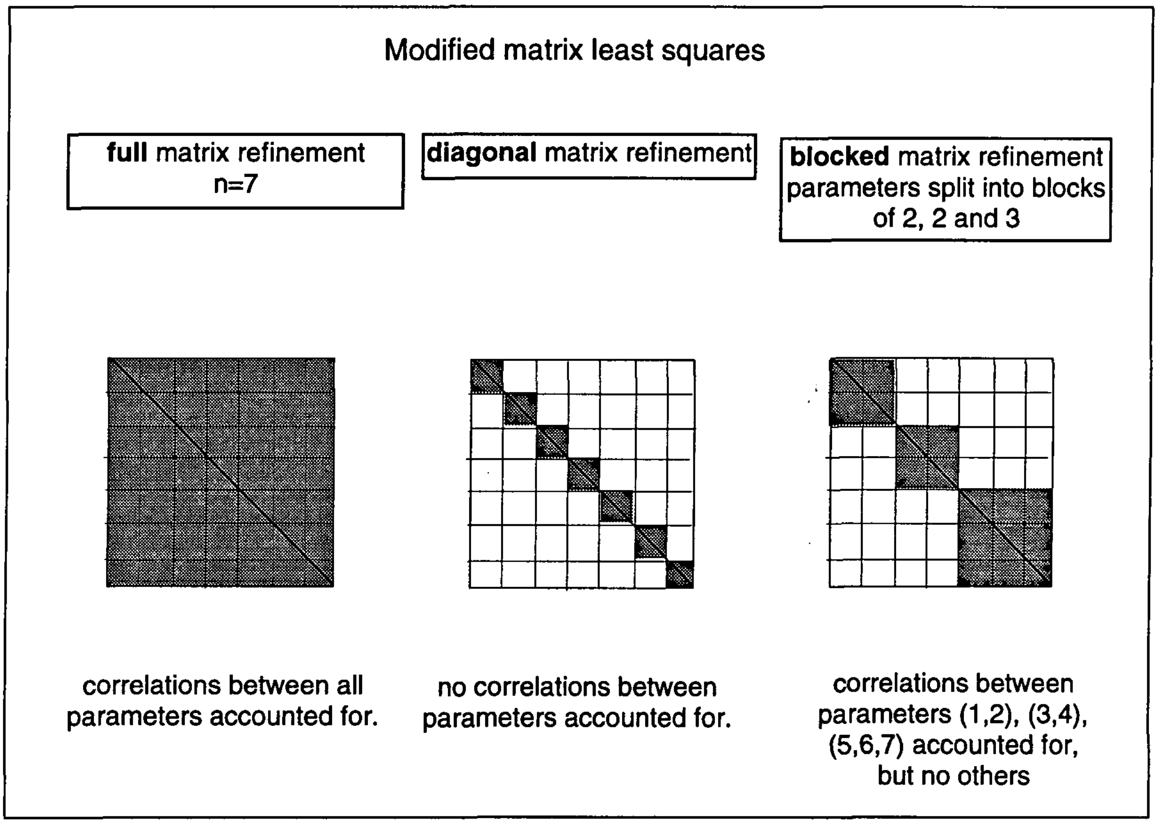
Although there is a lack of high resolution information in the diffraction pattern, there exists a store of structural knowledge to help compensate. A library of stereochemical information has been built up from sources including amino acid and oligopeptide structures and spectroscopic measurements. This information can be introduced into a refinement by the application of restraints or constraints.

A restraint restricts a model property to a realistic range of possibilities, weighted by the reciprocal variance of the property, whereas a constraint confines it to a specific value. If the weight on a restraint is infinite, it becomes a constraint. In their application, both increase the effective data to parameter ratio, in different ways. A constraint is an exact mathematical condition leading to the elimination of one or more least-squares parameter because they can be expressed exactly in terms of other parameters. A restraint is applied in the form of an additional weighted observation, with the discrepancy between model and ideal values to be minimised in the same way as the structure factor residuals (11). All observational terms contribute to a grand minimisation function (12).

$$\varphi(x) = \Sigma_i 1/\sigma_{x,i}^2 \, (g_{x,i}^{\,\mathrm{ideal}} - g_{x,i}^{\,\mathrm{model}})^2 \tag{11}$$

$$\Phi = \Sigma_x \varphi(x) \tag{12}$$

The storage space and time required for a refinement can be greatly restricted by considering only the diagonal terms of the normal matrix. This is a very rough approximation since it ignores all correlation between parameters. Block diagonal refinement involves the division of parameters into subsets, and contruction of a sparse matrix, with correlations within a subset accounted for, while those outside are ignored. This is a good compromise, especially if the blocks are different for each cycle of refinement.


Modified matrix least squares
full matrix refinement n=7
diagonal matrix refinement
blocked matrix refinement parameters split into blocks of 2, 2 and 3
correlations between all parameters accounted for.
no correlations between parameters accounted for.
correlations between parameters (1,2), (3,4), (5,6,7) accounted for, but no others

Conjugate gradient refinement is an approach which is often adopted as an acceptable compromise for macromolecular refinement. The method involves the storage of the parameter shifts between cycles. During each cycle, the direction vector of parameter shifts is modified by a factor derived from the direction vector for the previous cycle. The factor is larger when a parameter appears to creep in the same direction in successive cycles. This can be thought of as a method of approximating the second derivative of the function. It helps prevent the refinement becoming trapped in a false minimum and accelerates convergence while limiting time and memory consumed since the structure factor derivatives are computed only once, although the geometric information is updated on every cycle. Convergence properties are less good than for full matrix least-squares refinement. An additional disadvantage of the conjugate gradient method is that a reliable estimate of errors is not obtained.

PROLSQ (Hendrickson, 1985, Konnert & Hendrickson, 1980) is a least-squares refinement package developed to deal with the problems posed by the refinement of protein crystal structures outlined above. Minimisation of structure factor residuals is performed (7) using a conjugate gradient algorithm. Co-ordinates and individual isotropic thermal parameters may be refined for each atom, although, for lower resolution data and in early stages of refinement, one overall B factor derived from the Wilson plot (Wilson, 1942; Chapter 5A) can be fitted. A system of restraints is imposed during the refinement. The relative weight of observed data and restraints can be adjusted. The weight given to the restraints should be reduced at later stages in the refinement to allow maximum exploitation of the information in the diffraction data.

The system of stereochemical restraints implemented in PROLSQ is based on values in the Engh & Huber (1991) set. Restraints are applied to bonding distances and angles, to impose the planarity of groups and chirality at asymmetric centres. Non-bonding contacts shorter than given minimum values have a repulsive restraint imposed. Selected torsion angles, sidechain rotamers and peptide bond torsions are restrained to the ideal values to which model values correspond most closely.

Restraints on atomic thermal parameters follow the rigid bond principle. The root mean square displacements of atoms within a protein structure are an order of magnitude larger than the distortions observed in chemical bonds. This implies highly correlated atomic motion. B factors of bonded atoms are thus restrained to similar magnitudes.

Restraints vary in strength of application. Bond distance restraints are strong, since the ideal bond distances are known to high accuracy for naturally occurring amino acids and the observed variances are small. The distributions of torsion angles are wider, so torsion angle restraints are softer.

Non-crystallographic symmetry (NCS) can be exploited in the imposition of two types of restraints. NCS related objects are superimposed, allowing an average structure to be defined. Positional restraints on deviations from this average structure can then be applied, also restraints imposing similarity of equivalent temperature factors. These NCS restraints can confer stability at early stages in a refinement before being relaxed to allow genuine differences between the NCS related species to emerge.

PROLSQ also imposes limits resisting excessive shifts in parameters in a single cycle. This can be a problem in badly defined parts of the structure and can cause instability in a refinement.

Technological progress in X-ray crystallography in recent years includes the construction of high intensity X-ray sources such as dedicated and purpose built synchrotrons, the development of area detectors with rapid scanning and readout properties and cryogenic cooling techniques which allow crystals to be preserved at liquid nitrogen temperature during and between data collections. These achievements have facilitated the collection of an increasing number of high quality, atomic resolution data sets on protein crystals.

An atomic resolution data set has been defined by Sheldrick (1990) as one in which data extend to at least 1.2 Å resolution, at least 50 % of the data in the outer shell possessing intensities > 2σ. There have also been vast expansions in the storage and operational capacity of computers. As a result of these advances, methods for both solution and refinement of crystal structures which were previously limited in application to small molecules have increasing relevance in the field of high resolution protein crystallography.

The SHELXL refinement package (Sheldrick & Schneider, 1996) was initially developed for small molecular refinement but is now evolving to meet the demands of refinement of molecules of increasing size. SHELXL refinement allows greater flexibility in the treatment of higher resolution protein structures where data quality makes this justifiable.

SHELXL can be used for refinement with a choice of minimisation algorithms: conjugate gradient, blocked and full matrix refinement. At the end of the refinement, an error assessment can be made by running some final cycles using the full matrix.

SHELXL refines against intensity data rather than amplitudes. This is especially advantageous for macromolecular data, particularly those with pseudosymmetry, since they include many weak reflections. During data collection, some weak reflections may be recorded with negative intensities. If the intensity data are then converted to structure factors, these weak reflections will be ascribed zero amplitudes. It is optimal to use all observations in refinement, with suitable weighting, rather than employ an intensity cutoff. This is achievable with the employment of F^2 refinement.

In SHELXL atoms can be modelled with either individual isotropic thermal parameters or, if the data quality permits, with anisotropic thermal tensors. This more accurate modelling of static and dynamic displacements within a structure reduces the rms electron density in the (F_o-F_c,α_c) synthesis, allowing features such as solvent sites and alternate side chain conformations to become apparent above the background noise. The modelling of alternate conformations and partially occupied atoms, with the refinement of occupancies, is also facilitated within SHELXL.

A large and flexible range of restraints are available which can be tightened when modelling diffuse regions of the structure and relaxed where they are not necessary. Geometrical restraints can be applied to interatomic distances, molecular planes and chiral centres. Chemically but not crystallographically equivalent distances can be restrained to be equal, which is useful where reliable dictionary values are not available. Two types of restraint are applied to anisotropic thermal parameters, a rigid bond restraint similar to that employed in PROLSQ can be applied to the component of the ellipsoid along the bond, while a weaker similarity restraint can be applied to transverse components.

Anisotropically refined solvent atoms can be restrained so that the axes of the thermal ellipsoid remain approximately equal. This “isotropic” restraint helps prevent the atoms from becoming non-positive definite, the volume of the thermal ellipsoid refining to a negative value. This can easily happen because an ellipsoid is an inadequate model for the density of a typical solvent site. Besides smearing of the density due to thermal motion in a similar fashion to the libration of atoms at the ends of long side chains, such as Lys and Glu, the density at a ‘solvent site’ may result from the overlap of several closely situated, partially occupied sites, these formations often having the appearance of a chain of beads.

As the model becomes increasingly sophisticated, it is important to differentiate between modifications justified by observation and overfitting of the model. The R_{free} test (Brünger, 1993) can be applied as an indicator. A small percentage of reflections is removed at random from the working data before any changes are made to the model. The refinement is therefore not biased by this subset of data. R_{free} is defined as the R factor (14) evaluated using the omitted data. If the drop in R_{free} on the introduction of a change, such as movement from isotropic to anisotropic modelling of thermal parameters, is of the same magnitude as that in the R factor derived from the working data set, this is an independent verification of the protocol. However, the R_{free} test is insensitive to adjustments made in the final stages of the refinement of an atomic resolution structure, such as the modelling of alternate solvent networks and no validation tool can compensate for lack of chemical knowledge and experience. Another problem with the evaluation of R_{free} is that the absence of a percentage of data from the refinement is not optimal, since the electron density we are attempting to model is a Fourier transform of the entire diffraction pattern. Missing reflections result in ripples in density maps, which can be misleadingly localised and lead to modelling errors.

$$R = \Sigma \mid |F_o| - |F_c| \mid / \Sigma |F_o| \qquad (14)$$

A further fundamental difference between small molecule crystals and those of macromolecules requiring serious consideration during refinement lies in the solvent structure. The treatment of solvent is discussed in detail in Chapter 5B.

Protocol used for protein G and rubredoxin refinements

This chapter is concerned with the refinement of atomic resolution data sets for two small proteins, using essentially the same protocol.

In both cases the initial model for the refinement was derived from an already existing isotropic model of the structure. 5% of the reflections were separated from the working data set prior to refinement for the evaluation of R_{free}. The co-ordinates were randomised by rms 0.3 Å and a number of cycles of isotropic refinement were performed in an attempt to remove the memory of these reflections from the model. Isotropic refinement using PROLSQ was followed by anisotropic refinement using SHELXL-93 (Sheldrick, 1993). As already mentioned, the validity of the anisotropic model can be assessed by observing the change in R_{free} following its introduction. However, at atomic resolution the use of an anisotropic model is always justified (Dauter *et al.*, 1995). Hydrogen atom positions were not refined, but calculated using a riding model. During the course of refinement an Automated Refinement Procedure (ARP, Lamzin & Wilson, 1993) was employed, as explained below, for the construction and modification of the solvent network only, a procedure designated ‘restrained ARP’. Thus, the solvent structure underwent modification in real space in between rounds of least-squares refinement. Fourier maps were calculated using the CCP4 suite of programs including FFT (Ten Eyck, 1973).

Stereochemical restraints applied during both PROLSQ and SHELXL-93 stages of the refinement were based on values from the Engh & Huber (1991) set. The restraints applied are listed in Table 1.

Table 1. Stereochemical Restraints applied during refinement

restraints applied to:	*PROLSQ σ*[a]	*SHELXL-93 σ*[a]
bonding distances		
1-2, 1-3, 1-4, H-bond/metal co-ordination	*0.02, 0.04*[b]*, 0.05, 0.05 Å*	*0.03 (0.01)*[c]*, 0.03 (0.01)*[c]*, -, - Å*
thermal parameters		
similarity main chain 1-2.1-3,side chain 1-2,1-3	*4, 5, 6 ,8* $\AA^2$	*4* $\AA^2$ *(0.05* $\AA^2$*)*[d]
rigid bond restraint - component. along bond direction	-	*0.8* $\AA^2$ *(0.01* $\AA^2$*)*[d]
planarity, chiral volume	*0.02 Å*[e]*, 0.15* $\AA^3$	*0.2* $\AA^3$*, 0.2* $\AA^3$
non-bonding contacts		
single torsion	*0.3 Å, Dinc.= -0.3*[f]	-
multiple torsion	*0.3 Å, Dinc.= 0*[f]	
H-bond X..Y	*0.3 Å, Dinc.= -0.2*[f]	
H-bond X-H..Y	*0.3 Å, Dinc.= -0.9*[f]	
torsion angles		
planar (eg peptide Ω)	*3 °*	*0.3* $\AA^3$ [g]
staggered (eg aliphatic χ)	*15 °*	
transverse (eg aromatic χ)	*20 °*	
solvent		
isotropic restraint: similarity thermal ellipsoid axes	-	*8* $\AA^2$ *(0.1* $\AA^2$*)*[d]

[a] σ of ideal distribution (not exactly for SHELX)

[b] 0.03 Å for protein G

[c] restraints applied to bond distances involving multiple conformations

[d] in terms of rms atomic displacement, $<u>^2 = B / 8\pi^2$

[e] distance of deviation from plane

[f] allowed separation = sum of Van der Waal's radii + Dinc

[g] chiral volume restraint on peptide bond carbonyl carbon

Real space refinement matches the expected and actual shape of density around an atom on the ($3F_o$-$2F_c$,α_c) map and moves the atom to improve its sphericity. This corrects the positioning of diffuse, possibly partially occupied, atoms which otherwise tend to drift towards the edges of the density.

Calculated structure factors were scaled prior to map calculation using a procedure which applies a correction accounting for the diffuse solvent scattering contribution, which is crucial for the fitting of data at 5 Å and lower resolution. The modelling of diffuse solvent is considered in Chapter 5B. Another possible improvement in this real space refinement, the use of sharpened density maps, is discussed in chapter 5A.

Estimates of errors in bond distances, angles and refined occupancies were obtained by running a final cycle of SHELXL-93 least-squares refinement with a matrix of overlapping blocks of parameters, using the final model and the entire set of data. All restraints were removed and the refinement was prevented from crashing by setting the allowed shift on all parameters except the overall scale factor to zero. The accuracy of these estimates is limited, since correlations between parameters not refined in the same block remain unaccounted for. Ideally, the full least-squares matrix should be constructed.

Chapter 4A:

Anisotropic refinement of the protein G IgG binding domain III

Introduction

What is an Immunoglobulin?

Immunoglobulins, or antibodies, are soluble proteins produced by plasma cells constituting the recognition element of the humoral immune response system. They are synthesised in response to the presence of a foreign substance, an antigen, which may be a protein, polysaccharide, nucleic acid, synthetic peptide or small molecule synthetically attached to a protein (hapten). Each antibody-producing cell has the capability of producing one sort of antibody only.

Polyclonal antibodies to a specific molecule may be obtained by injecting the substance into an animal such as a rabbit. The rabbit will develop antibodies, which may be extracted in antiserum. The antibodies obtained will be heterogeneous, varying in affinity and specificity for the antigen, since they were produced by many populations of cells. The investigation of the structure and function of immunoglobulins was greatly advanced by the development of a method for production of homogenous monoclonal antibodies. A single antibody producing cell is fused with a myeloma tumour cell. This then divides rapidly, producing hybridoma cells, which all possess the ability to produce the same specific antibody.

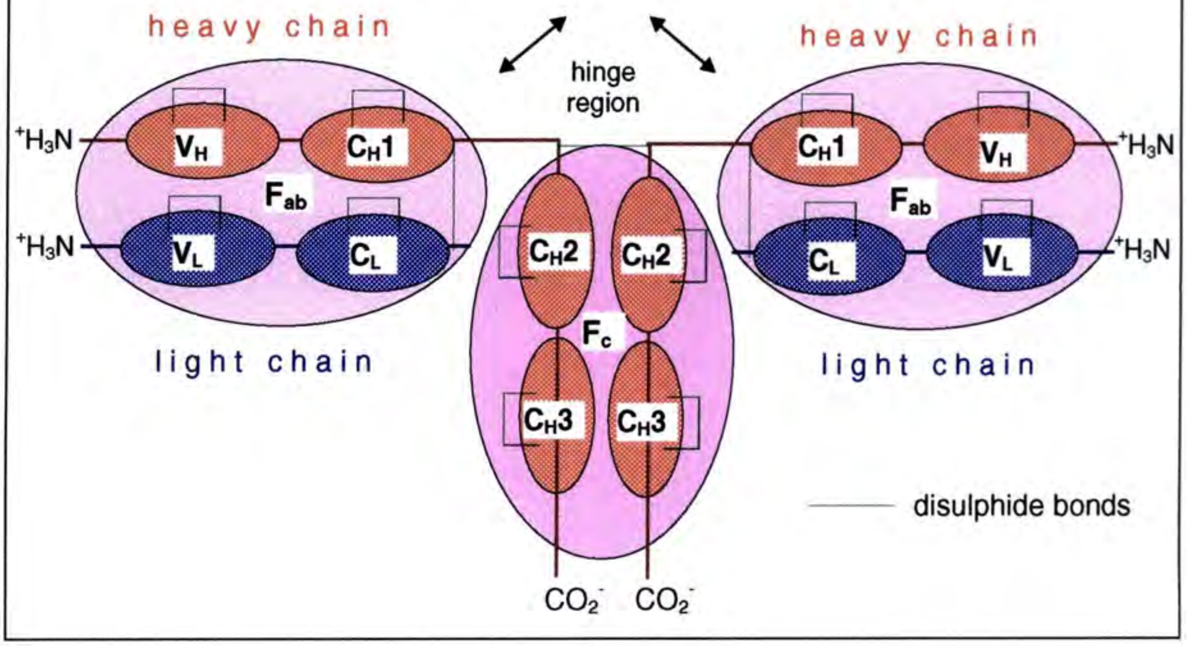
Immunoglobulin G is the principle class of antibody present in the blood plasma. The structure of IgG is schematically depicted in Figure 1. It has a moelcular mass of around 150 kDa. When cleaved with papain, three 50 kDa domains are obtained. Two are identical antigen-binding domains, Fab. The other, known as Fc, is responsible for effector functions, the processes by which the antigen-antibody complex is dismembered. The three domains fit together in a Y shape, with the Fc constituting the stem. Ig is made up of four polypeptide chains, two identical 25 kDa "light" chains and two identical 50 kDa "heavy" chains. The C-terminal ends of the heavy chains constitute the Fc domain. Each Fab domain consists of a light chain and the N-terminal part of a heavy chain. The chains are held together by disulphide linkages. The antigen-binding site is a cleft located at the N-terminal end of the Fab domain, at the top of the "Y". The link between the domains is flexible, allowing hinge-like movement of the two Fab segments, thus the distance between the antigen binding sites can be varied to optimise binding to antigens with multiple recognition sites, for example, viruses.

Both heavy and light chains consist of a variable N-terminal region and a constant region. Within the variable regions are hypervariable segments which form the antigen-binding site. The Fab and Fc fragments both contain four globular subdomains arranged in a tetrahedral manner; Fab: V_L, C_L, V_H & C_H1 and Fc: $2C_H2$ & $2C_H3$. Each subdomain adopts the structural motif known as the Immunoglobulin fold, consisting of a pair of β-sheets; three-stranded and four-stranded, back-to back, bridged by a disulphide bond.

Two types of light chain exist, κ and λ, with around 40% sequence identity in the C_L subdomains. The heavy chain varies with the class of antibody. Some classes have larger Fc domains and exist as larger structures, with multiple copies of the (Fc 2Fab) unit. Membrane-bound immunoglobulins are also produced, which are embedded in the membrane via a hydrophobic extension to the Fc domain.

The specificity of the antibody for the antigen results from the fact that the binding site is composed of chain segments with sequences unique to each antibody. It is not caused by conformational changes following the invasion of antigens into the environment and novel antibodies are not synthesised in response to the presence of an antigen. The capacity to synthesise each type of antibody is inherent in the system.

Figure 1. Schematic illustration of the structure of IgG



What is Protein G?

Protein G is a large multi-domain cell surface protein of group G *Streptococcus* which binds to Immunoglobulin G. Many pathogens possess such proteins, which assist in the evasion of the host's immune response by binding to host antibodies, mimicking self markers commonly displayed by host cells (Boyle, 1990). Other examples are protein A from *Staphylococcus aureus* and protein L from *Peptostreptococcus magnus*. Ig binding proteins typically possess two or more Ig-binding domains, each comprising 50-60 residues with close to 100% identity between them. These subdomains retain activity when individually expressed.

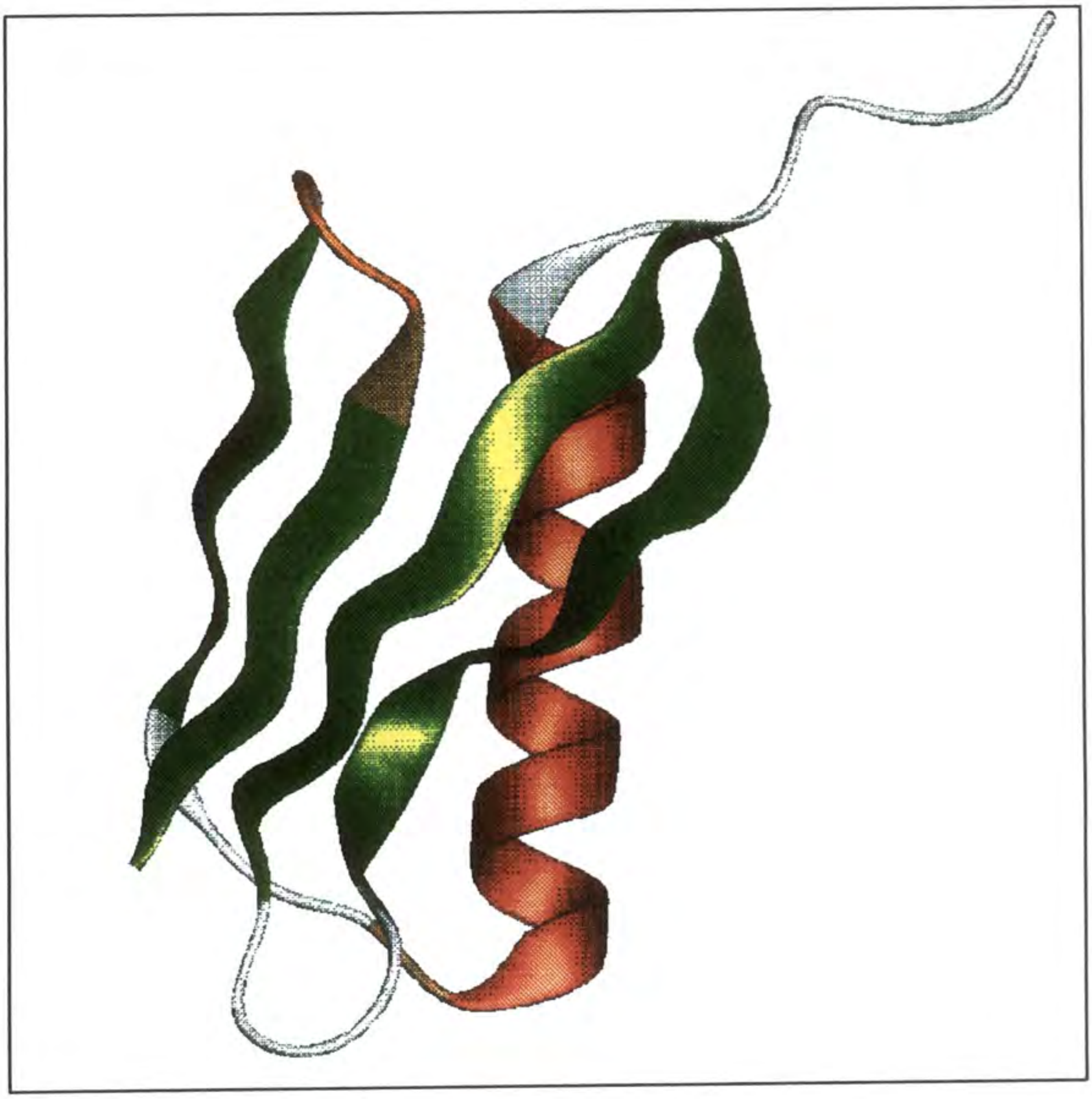


Figure 2. Overall fold of protein G IgG binding domain.

Beta strands are shown in green and the alpha helix in red.

Drawn with QUANTA.

The role of these Ig-binding proteins in enhancing microbial virulence makes them a target for the understanding and combating of infection. The common clinical symptom caused by group G *Streptococcus* is pharyngitis, but it can be responsible for more serious infections, such as septicaemia. The Ig-binding affinity of these proteins is high, comparable in strength to antibody-antigen binding and their specificity is low. Typically Igs of several species and classes are bound by the same protein. The nature of the protein-protein interactions is of interest. Affinity for Ig is unlikely to be affected by the binding of antigens, which has led to these molecules becoming crucial tools in the screening and purification of antibodies, proteins generally and other biomolecules in immunology and molecular biology.

Protein G contains three IgG-binding domains. There is a high degree of sequence similarity between published sequences of protein G IgG-binding domains, suggesting that strict sequence conservation is necessary for the retention of binding ability. The secondary structure of protein G Ig-binding domains comprises a central α-helix, packed against a four-stranded antiparallel β-sheet: -1, +3x, -1 topology, after the notation of Richardson (1981), Figure 2. The domains possess exceptional chemical and physical stability (Achari *et al.*, 1992). The B1 domain has a melting temperature of 87° C and is not denatured in 8 M urea. One factor contributing to this stability is that around 95% of the residues are incorporated in secondary structure, ensuring a large number of stabilising hydrogen bonds. A typical value for a domain of this size is 75% (Gronenborn *et al.*, 1991). The overall fold of the molecule enhances its stability, with the two ends of the chain making up the central strands of the β-sheet and a tightly packed, hydrophobic core, coupled with a very hydrophilic exterior. This results in a large negative solvation-free energy of folding, -55 ± 2 kcal mol^{-1} for B1.

Protein G binds to a wide range of mammalian IgGs, but not to other classes of Ig (Guss *et al.*, 1986). The binding domain utilises two almost completely non-overlapping sets of residues on its surface to recognise two separate sites on the Ig (Lian *et al.*, 1994), one on the Fc fragment, one on the Fab domain. The affinity of Protein G for IgG binding sites varies between IgG species. For human IgGs, interaction occurs preferentially to the Fc domain, while for mouse Ig$_1$, binding is with the Fab domain.

Complex formation with the Fab domain of the antibody occurs through the antiparallel interaction of the second β-strand of the protein G domain with the last β-strand of the C$_H$1 subdomain, to form an extended β-sheet across the two domains. A minor binding site, comprising residues from the C-terminus of the α-helix and the first β-strand from C$_H$1, packs against the β-β contact region, forming a hydrophobic interior to the interaction site. The crystal structure of Protein G domain III, complexed with a Fab fragment from mouse IgG$_1$ (Derrick & Wigley, 1994), gives insight into the manner in which high affinity, low specificity binding to Ig is achieved. NMR studies on a complex of domain II with a Fab fragment from mouse IgG$_1$ (Lian *et al.*, 1994) support the fact that this interaction persists in solution and is not an artefact of crystallisation. Protein G binds to the least variable part of the Fab domain. Out of a total of twelve hydrogen bonds, eight connect main chain atoms in C$_H$1, including the five involved in the β-β contact. In addition, the C$_H$1 residues with side chains involved in the interaction are highly conserved between different subclasses and species.

Binding to the Ig Fc domain occurs in a cleft between C_H2 and C_H3 subdomains. The binding site incorporates residues on the α-helix and the third β-strand of the protein G domain (Gronenborn & Clore, 1993).

Protein A binds to mammalian IgG Fc fragments in a complementary fashion to Protein G (Guss *et al.*, 1986). The two proteins exhibit competitive binding to Fc domains, indicating that the binding sites are close or overlapping (Lian *et al.*, 1991). However, no significant homology has been found between protein A and G IgG-binding domains and the secondary structures are not alike. The protein A solution structure consists of three helices (Torigoe *et al.*, 1990; Gouda *et al.*, 1992), one of which is disrupted in the crystal structure of the protein A/ Fc complex (Deisenhofer, 1981). There is evidence that the protein G recognition of Fc principally involves the helix, in a similar fashion to the protein A - Fc interaction. It is possible to superimpose one of the two Fc interacting helices of protein A on to the G helix. This results in the alignment, with respect to Fc, of the third β-sheet of protein G in a similar position to that of the second interacting helix of protein A (Gronenborn & Clore, 1993). The manner in which these two different structures from bacterial proteins recognise the same site on a host Ig is a good example of convergent evolution.

The protein L IgG binding domains adopt the same fold as the protein G chains. However no significant homology could be found (Kastern *et al.*, 1992). Protein L binds to the Fab domains of all IgG classes. It interacts with the framework of the variable domain of κ light chains and has low affinity for λ chains. A hybrid molecule, protein LG, containing Ig-binding domains from L and G, has been found to bind to a very wide range of Igs. (Kihlberg et al., 1992). Thus, proteins A and G possess similar functions but different structures. Conversely, proteins L and G adopt similar structures while performing different functions.

Summary

Derrick & Wigley (1994) grew crystals of the third IgG binding domain (domain III) of protein G, collected X-ray diffraction data to a resolution of 1.1 Å, obtained a structure solution, performed refinement and deposited the resulting co-ordinates in the PDB, code 1IGD. An outline of this study, which preceded the work performed as a part of this PhD, follows. In the present study, the data have been reprocessed and anisotropic refinement has been carried out, with the aim of assessing the additional information available from the anisotropic model.

Previous experiment: (Derrick & Wigley, 1994)

Protein and Crystallisation

Expression and purification of domain III was carried out as described by Lian *et al.* (1992). Crystals were grown by hanging-drop vapour diffusion, with a reservoir solution containing 10 mM sodium acetate buffer (pH 4.8), 0.01% sodium azide and the precipitant, 24-26% PEG 4000. Crystals of dimensions up to 0.75 x 0.50 x 0.50 mm grew in one week. The crystal properties are summarised in Table 1.

Table 1. Crystal properties, data processing statistics

Crystal symmetry:	Orthorhombic, $P2_12_12_1$	Molecules per asymmetric unit; 1
	Previous Refinement	**New Refinement**
Cell dimensions (Å)	a= 34.9, b=40.4, c=42.2	a= 34.78, b=40.28, c=42.19
V_m (Å^3 da^{-1})	2.29	2.23
Resolution (Å)	25 - 1.1	10 - 1.1
R_{merge} (%)	5.8	3.7
N_{unique}	23 530	24 145
Completeness (%)	95.0	97.9

Data Collection and Processing

Data were collected at room temperature on the X31 beamline at the EMBL outstation using monochromatic radiation of 0.72 Å wavelength and a Hendrix-Lentfer Image plate detector, the prototype MAR scanner. For the most effective solution and refinement of atomic resolution structures, data must be as complete as possible. The very low resolution data play a crucial role in refinement, e.g. (Dodson *et al.*, 1996). In order to collect data over a wide range of resolutions on a MAR imaging plate at a synchrotron source, it is necessary to collect several data sets.

To maximise the resolution which can be collected, crystal to detector distance must be short, a minimum of 80 mm is possible on beamlines at EMBL Hamburg. Due to the limits of its dynamic range, strong, low resolution reflections overload the detector when exposure time is sufficiently long to allow reliable measurement of weak, high resolution intensities. Therefore, data are collected in several runs with the crystal to detector distance increased and exposure time successively reduced between each one. Four data sets were collected (at distances 140, 190, 290 & 450 mm) with resolution ranges, 1.8 - 1.1 Å, 4.0 - 1.4 Å, 10 - 2 Å, 25.0 - 4.0 Å, the low resolution limit for an image being determined by the radius of shadow cast by the beam stop at the centre of the image plate. The data were processed using the CCP4 (1979) suite of programs, giving statistics shown in Table 1.

Structure Solution and Refinement

Structure solution by molecular replacement, using structures obtained by NMR spectroscopy as search models, proved unsuccessful, so heavy-atom isomorphous replacement was required. Structure solution was achieved using data from a single lead acetate derivative. An initial model was built using the interactive graphics program FRODO (Jones, 1978), comprising residues 2-13, 16-50 & 55-61, with numbering from the N-terminal Met, using the sequence described by Lian *et al.* (1992). Following a few cycles of restrained least-squares refinement (Hendrickson & Konnert, 1980), construction of the remaining loops was possible, using a ($2F_o$-F_c) map. Alternate cycles of model rebuilding and least-squares refinement with phase extension followed. There was also a simulated annealing step using X-PLOR (Brünger *et al.*, 1990). Characteristics of the final (new) model are summarised in Table 2 in comparison with the 1IGD (old) model.

Table 2. Comparison of new and previously refined models

	old		new
R (%)	19.3		9.4
total water	120		130
B protein, mean ($Å^2$)	11.9		12.6
B water, mean ($Å^2$)	40.2		41.2
rmsd, CA atoms(Å)		0.054	
rmsd, all protein atoms (Å)		0.423	

Anisotropic Refinement

The work undertaken as part of this PhD commenced at this point.

Data

The data, collected as described above, were reprocessed using DENZO (Otwinowski & Minor, 1993). This was deemed to be necessary following initial test refinement runs, which suffered from problems including inexplicable program crashes, fluctuations in R factor and a high level of noise in density maps. Problems in scaling of the low resolution data were traced to merging errors in the medium resolution data set, (10-2 Å) resulting from the fact that the beam-stop shadow was too large in this set of images and also off-centred. This caused a select set of data to be unobserved. Once this problem was accounted for, the merging of the low resolution data improved. Reprocessing of the data gave slightly different cell parameters (Table 1) with new data processing statistics. Errors in the very low resolution intensities remained large, so a resolution cutoff of 10 Å was applied. Refinement was performed using a working set of 22944 data - 95% of the total. The remaining 5%, removed at random, were utilised in the calculation of R_{free}, a cross-validation method of assessing the progress of the refinement (Brünger, 1993).

Model

The starting point for the refinement was the previously determined isotropic model, PDB code 1IGD. Cell parameters were changed to the values obtained during the reprocessing of the data. All solvent atoms were removed. A random positional error with rms 0.3 Å was introduced into the model to assist in the eradication of the memory of individual reflections from the model.

Refinement

Restrained least-squares refinement of atomic positions and thermal parameters was performed, in the initial stages using PROLSQ and then using SHELXL-93 (Sheldrick, 1993). Stereochemical restraints to parameters taken from the Engh & Huber (1991) set were applied during both stages of the refinement, as explained in the introduction to this chapter. Hydrogen atom positions were not refined, but calculated using a riding model. ARP was employed for modification of the solvent structure in real space. Visualisation and manual rebuilding of the model were performed using FRODO.

The course of the refinement is summarised in Table 3. The first stage consisted of PROLSQ refinement alone. This was a precaution to help remove memory in the model of the reflections in the R_{free} data. Further isotropic refinement followed, during which building of the solvent network was commenced. A cycle of ARP was run after each PROLSQ cycle, with ARP set to modify solvent atoms only. At first, the limits for addition and removal of solvent atoms were set to 20 per cycle, allowing the solvent network to be constructed. The numbers of atoms added and removed per cycle were then reduced to 5, allowing more stable refinement of the existing solvent positions. When this refinement had converged and no more solvent was being added, it was judged time to start anisotropic refinement.

Table 3. The course of refinement

Cycles	Stage		R (%)	*ΔR(%)*	R_{free}·(%)	*ΔR_{free}(%)*	water sites	$\rho_{rms,\,F}$ #	$\rho_{rms,\,\Delta F}$ ##
	-	randomisation of co-ordinates from previous refinement	35.9	-	37.3	-	0		
1-15	1	isotropic refinement, PROLSQ	26.6	*-9.3*	28.3	*-9.0*	0	0.557	0.093
16-30	2	PROLSQ + real space refinement & construction of solvent network, ARP	17.5	*-9.1*	19.4	*-8.9*	118	0.562	0.090
31-40	3	anisotropic refinement , SHELXL-93, ARP	10.1	*-7.4*	12.9	*-6.5*	119	0.500	0.049
41-50	4a	introduction of 4 double conformations, occupancies of 3 other residues refined	9.5	*-0.6*	12.8	*-0.1*	121	0.493	0.047
51-60	4b	8 double conformations & Lys 15 occupancy refined	9.5	*0*	13.0	*+0.2*	123	0.495	0.047
61-70	5	sharpened ($3F_o$-$2F_c$) maps input to ARP; ($F^{0.2}E^{0.8}$)	9.4	*-0.1*	12.5	*-0.5*	130	0.491	0.046
71-80	6	100% data used (unsharpened maps)	9.4	*0*	(9.5)*	*(-3.0)**	131	0.504	0.047

* the R_{free} data set is no longer independent of the refinement at this stage

# $\rho_{rms,\,F}$ = rms of ($3F_o$-$2F_c$) map density

## $\rho_{rms,\,\Delta F}$ = rms of (F_o - F_c) map density

Before anisotropic refinement commenced, five cycles of SHELXL-93 were run on the isotropic model. The resulting model had R and R_{free} 17.9% and 20.2%. This difference in R factors from the PROLSQ model principally results from the fact that SHELXL-93 refines against F^2 as opposed to F. Isotropic atoms remained isotropic for 1 cycle of refinement and were then refined anisotropically for 4 cycles. Atoms already anisotropic remain so throughout. After each round of SHELXL-93 refinement, ARP was run, adding and removing up to 5 atoms per cycle. R and R_{free} dropped to 10.9% and 15.0% in two rounds of refinement, demonstrating the validity of the anisotropic model and the protocol in general.

A prominent feature of a difference map, generated from a model with isotropic thermal parameters, is the presence of peaks close to atoms, especially in the more mobile parts of the structure, such as loop regions and at the ends of long side chains. Modelling of the anisotropy of the atoms reduces this difference density, lowering the rms density of the whole map and enabling parts of the structure which have been incorrectly or insufficiently modelled to become apparent from the remaining difference peaks. For this reason, manual fitting of the model to the density and the building of alternate conformations for disordered regions was left until several cycles of anisotropic refinement had been performed, to reduce the possibility of making incorrect alterations.

The fit of the model to the ($3F_o$-$2F_c$, α_c) and (F_o-F_c, α_c) density syntheses was inspected using FRODO. The side chain of Lys 24 was adjusted. The pattern of water sites around Asn 13 suggested that ND2 and OD1 should be interchanged. For four residues: Val 5, Glu 12, Val 26 & Glu 32, difference density clearly showed the presence of alternate conformations. Two alternative conformations of each side chain were modelled, with the combined occupancy of each pair constrained to 1.0. For three residues: Met 1, Lys 15 and Asp 40, difference density indicated some disorder and the correct side chain conformation was unclear. These residues were each modelled with a single side chain conformation and its occupancy was refined. Following further refinement, positions for second conformations of Met 1, Glu 20, Asn 40 and Asp 51 were modelled.

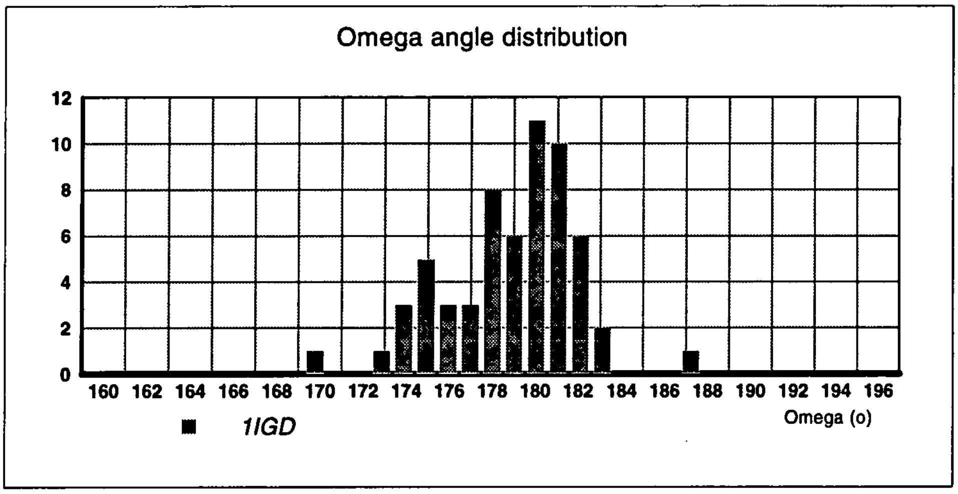
When manual adjustments to the model were deemed complete, a further refinement was carried out with the same protocol except that the $(3F_o\text{-}2F_c,\alpha_c)$ maps input to ARP were calculated using sharpened structure factors of the form $(E^{0.8}F^{0.2})$. This was an attempt to see if real space refinement using sharpened maps could improve the modelling of the solvent network, as explained in Chapter 5a. The working and R_{free} data sets were then combined and refinement was performed using 100% of the data, now without sharpening of the maps input to ARP. Lastly, a cycle of blocked least-squares refinement was run to provide estimates of errors in the final model, as described in the introduction. The estimated co-ordinate errors for each atom type were strongly correlated to the B factors. The mean values for main-chain CA, N and O were 0.0183 Å, 0.0143 Å and 0.0147 Å respectively.

Comparison of anisotropic model with 1IGD

Protein

A comparison of overall geometry of new and 1IGD models, Table 4, shows that mean bond distances are longer for 1IGD, with a mean difference in main chain bond lengths of +0.7%. The former refinement was performed using a pre-Engh & Huber dictionary. A second possible cause for the difference in bond lengths is the exclusion of hydrogen atoms throughout the 1IGD refinement. In the absence of hydrogen atoms, the unaccounted for electron density produces a tendency for bonds in the model to elongate.

Figure 3



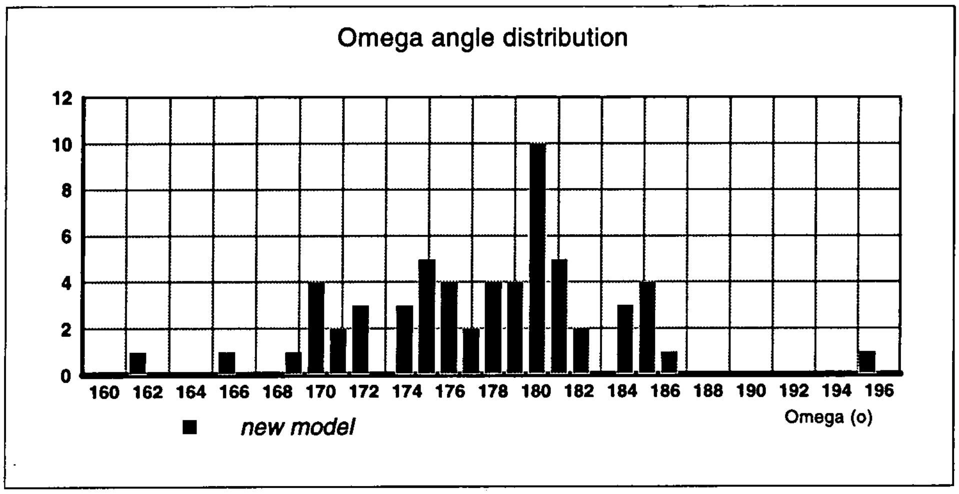


Table 4. Comparison of the geometry of the new and 1IGD models

	New model		target (Å)	**1IGD**	
	mean interatomic distance (Å)	σ interatomic distance (Å)		mean interatomic distance (Å)	σ interatomic distance (Å)
N-CA	1.448	0.015	1.458	1.468	0.011
CA-C	1.515	0.014	1.525	1.527	0.011
C-O	1.226	0.010	1.231	1.244	0.009
C-N	1.318	0.012	1.329	1.318	0.009
N-C	2.430	0.025	2.462	2.440	0.039
CA-O	2.383	0.021	2.401	2.400	0.021
O-N	2.232	0.016	2.250	2.262	0.015
C-CA	2.424	0.022	2.435	2.429	0.015
CA-N	2.414	0.020	2.425	2.417	0.026
	mean Ω (°)	σ[Ω] (°)	target (°)	mean Ω (°)	σ[Ω] (°)
	177.3	5.8	180.0	178.9	2.9

The distributions of distances for 1IGD are sharper than for the new refinement, Table 4. The distribution of the peptide torsion angle, Ω, Figure 3, is narrower, more peaked and centred closer to 180° in the former. This is a sign that, during the old refinement, too much weight was placed on geometric restraints. This forced the adoption of idealised stereochemistry, overriding information present in the X-ray data, thereby leading to loss of information in the model.

The furthest outliers to the new model Ω distribution are the angles before and after Trp 48, 195° and 162°. Figure 4 is an illustration of the possible consequences of the application of an over strong planarity restraint to this section of chain. Trp 48 is in the centre of the β3 strand, the surrounding chain interacting with β2 of the neighbouring molecule, O48 and N20 are H-bonded, and with β4 of the same molecule, via H-bonds N49-O58 and O47-N60. The bulky side-chain sits in a hydrophobic cleft between the β-sheet and the C-terminal end of the α-helix. Val 59 CB points towards the Trp ring and is separated from it by around 3.8 Å. A difference peak 2.5 σ in height lies between the Val CB and the Trp ring, at around 1.7 Å from the ring, in an otherwise very clean area of

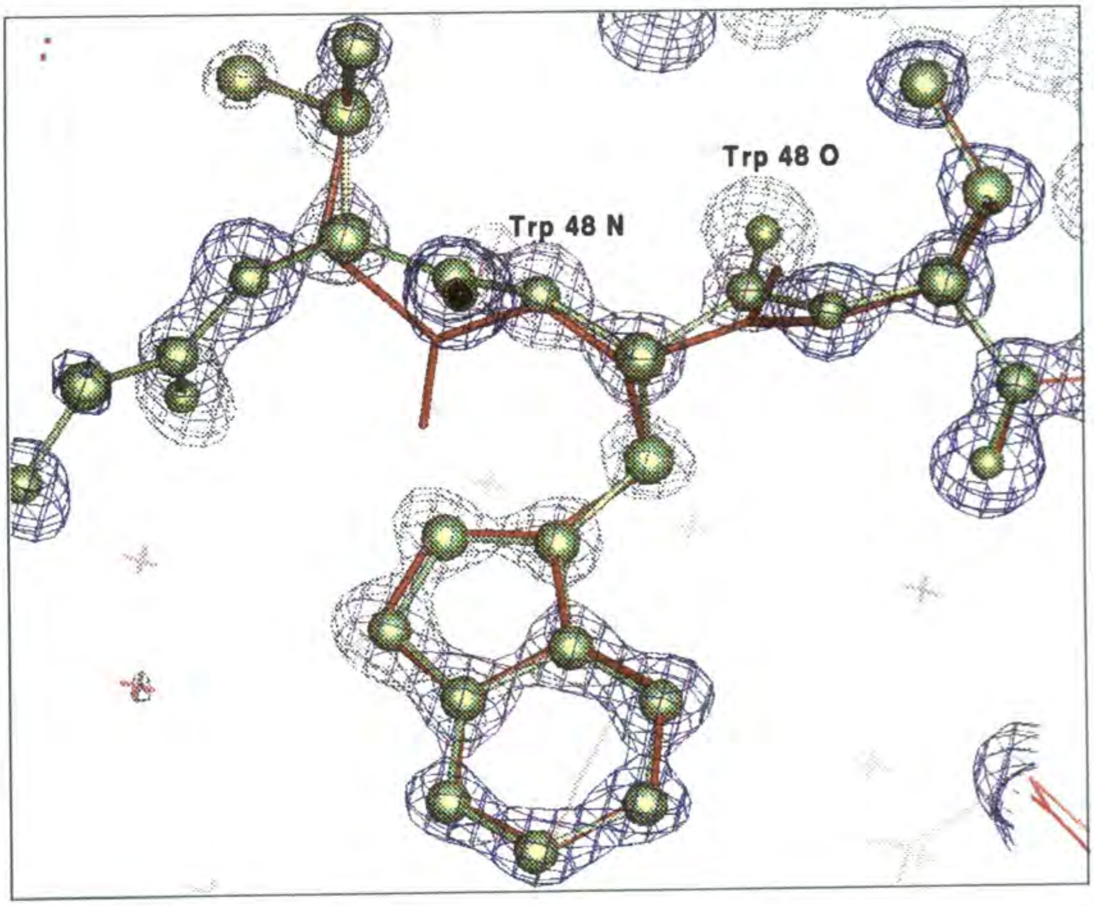


Figure 4:

The peptide torsions before and after Trp 48 were distorted from 165 and 192 to the 'ideal' value of 180 degrees. The geometry around these bonds was then re-regularised.
The green ball-and-stick model shows the original atomic positions. The distorted model is shown in red.
The fit of both models to the 3Fo - 2Fc density map is shown.

density.This suggests that there may be an alternate conformation of Trp 48, in which the ring is rotated by 180°,allowing one CG of an alternate conformation of Val 59 to lie in the density observed between the two residues.

While the mean B factor for solvent atoms in the new model is almost identical to the old value (see Figure 10), that for protein atoms increased by an average of +0.8 $Å^2$ per residue, Figure 5. This can be explained by the fact that refinement without application of a diffuse solvent correction, as in the case of 1IGD, results in a global underestimation of B factors, see Chapter 5B. There are nine residues for which the new refinement produced a lower mean B factor: Val 5, Ile 12, Lys 15, Glu 29, Glu 32, Lys 33, Glu 37, Asn 40 and Asp 45.

The change in map correlation (1) by residue, shown in Figure 6, exhibits the same trend as the change in B factors, with residues 5, 12, 15, 29, 32, 37, 40 and 45 all experiencing an improvement of over 10 % in their real space fit.

Correlation correlation coefficient, $k = (\langle xy\rangle - \langle x\rangle\langle y\rangle) / (\sqrt{(\langle x^2\rangle - \langle x\rangle^2)}\sqrt{(\langle y^2\rangle - \langle y\rangle^2)})$ (1)

All these residues fitted poorly into the density in the initial stages of the refinement. Two alternate conformations were modelled for the side chains of Val 5, Ile 12, Glu 32 and Asn 40. The density for Lys 15 CD, CE and NZ was weak and the occupancy of a single conformation was refined to a value of 60%. The density for Glu 29 and Lys 33 was also weak, but density for all atoms could be seen after refinement. Modelling of alternate side chains was attempted for Glu 37 and Asp 45, but was unsuccessful. Continued refinement greatly improved their fit and difference density around them was removed. It is likely that several of these residues were the victims of the noisy solvent interface, which resulted from incomplete modelling of the solvent in the early stages of refinement.

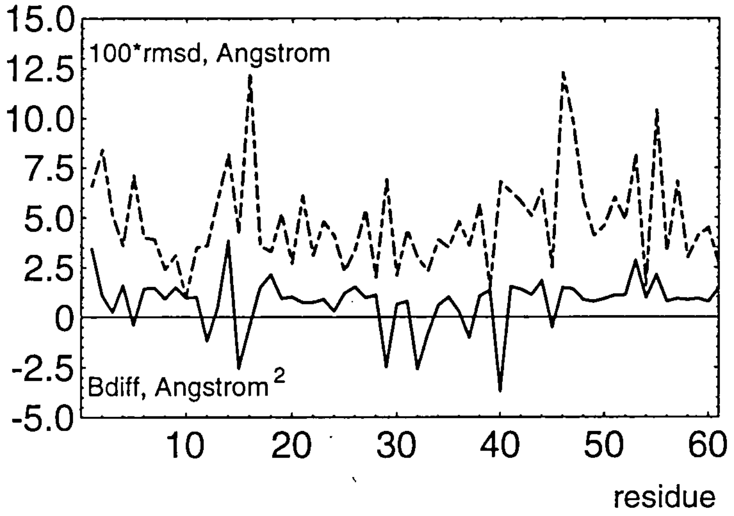


Figure 5.

Comparison of the new and 1IGD models by residue.

The dashed line shows the change in equivalent isotropic B factor:

<B(new)> - <B(1IGD)>

The continuous line depicts the rmsd between the CA atoms of the two models:

100 (rmsd (CA))

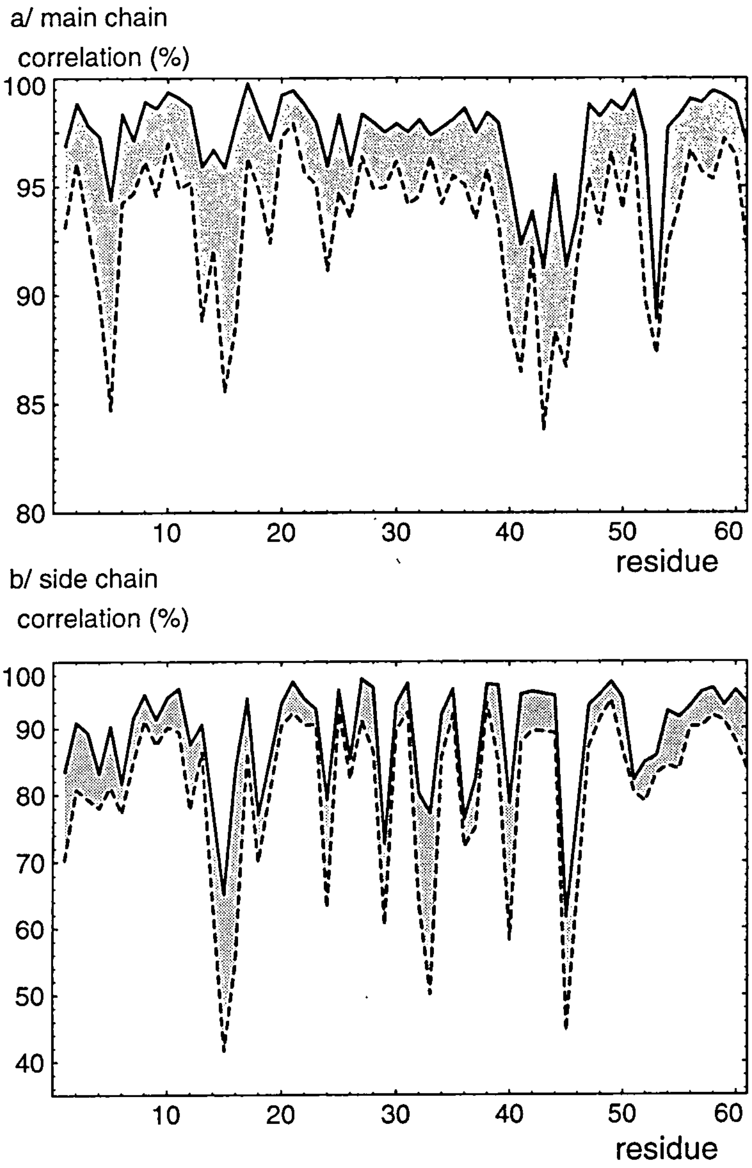


Figure 6.

Map correlation of the (3Fo-2Fc) map to the Fc map for the new and 1IGD models

a/ for main chain atoms

b/ for side chain atoms (excludes Gly)

k(new)(continuous line)and k(old) (dashed line) are plotted against residue number, where k is the correlation to the map per residue.

In the crystal structure of the protein G-F_{ab} complex (Derrick & Wigley, 1994) an antiparallel alignment of the second β-strand from protein G domain III, residues 15-22, with the 7th and last β-strand of the C_H1 domain, extends the 4-stranded β-sheet from domain III into C_H1. The end of the domain III α-helix and the 1st β-strand of C_H1 come together to form a secondary binding zone, which lies on top of the join between the two β-sheets, creating a hydrophobic binding region. NMR studies on a protein G domain II - F_{ab} complex (Lian *et al.*, 1994) show that the regions of the protein G domain involved in IgG binding in solution correspond to the binding residues in the crystal, so this alignment of the molecules is not merely an artefact of crystallisation. When protein G IgG binding domains are crystallised in the absence of IgG, a similar β-β interaction is commonly observed (Derrick & Wigley, 1994; Gallagher *et al.*, 1994): the antiparallel alignment of the second β-strand of one molecule with the third strand of the next, resulting in continuous ribbons of β-sheet extending through the structure.

Intermolecular contacts in the crystal have been examined, with particular attention to the β-β interaction region, Figure 7. Four main-chain hydrogen-bonding interactions were observed between β2 and β3, as listed in Table 5. In the new model, the placing of the Thr 16 side chain in a different rotamer, coupled with the modelling of two conformations for Asp 51 has lead to a new contact being established between OG 16 and OD1 51. The B factors of the main chain atoms involved in H bonding increased slightly, while those of the side chain atoms decreased. The lengths of already established contacts decreased by an average of 3.5% of the new mean distance, 2.83 Å. Other intermolecular hydrogen bonds were formed by very exposed residues, the N and C termini, NZ of Lys 9 and Lys 24 and the side chains of Glu 20, Asp 27, Gln 37 and Asp 41. The new refinement has resulted in improved modelling of these residues in diffuse density, as described above, allowing these contacts to be more clearly revealed.

Table 5. Comparison of intermolecular H-bonding in new and 1IGD models

(1)	(1) B_{new} - B_{old} ($Å^2$)	(2)	(2) B_{new} - B_{old} ($Å^2$)	d_{new} (Å)	d_{old} (Å)
β-β H-bonding contacts - main chain atoms					
O 16	-0.3	N 52	+1.4	2.95(1)	3.02
N 18	+1.6	O 50	+1.5	2.94(1)	3.01
O 18	+1.2	N 50	+1.5	2.88(1)	2.95
N 20	+1.0	O 48	+1.0	2.86(1)	2.92
new side chain β-β *H-bonding contact*					
OG 16	-10.2	OD1 51	-4.2	3.16(3)	> 4.5
other intermolecular H-bonds					
N 1	-0.9	OG1 60	+1.5	2.94(1)	3.00
N 1	-0.9	O 22	+1.8	2.94(1)	3.02
NZ 9	+5.3	OD1* 27	+1.5	2.84(2)	2.74
NZ 9	+5.3	OE2 29	+4.0	2.70(4)	2.89
NZ 24	-1.5	NE2 37	-3.5	3.27(4)	4.45
NZ 24	-1.5	OD2 41	.0.8	2.82(4)	3.21
NE2 37	-3.5	OD2 41	-0.8	2.82(3)	3.00
OE2a 20**	+2.2	OD1* 27	+1.5	2.96(1)	2.56
OE2b 20**	0	OD1* 27	+1.5	2.51(7)	2.56

* OD2 27 in 1IGD ** OE2 20 single site in 1IGD

WHATCHECK highlights the fact that the Asn 13 side chain in 1IGD should be rotated, as in the new model. Five unsatisfied potential hydrogen bonding atoms buried in the protein are listed for each model. Three of these listed only for the old model are Lys9 NZ, Lys15N and Glu 61 OE2. Asn13 ND2 features in both lists, although its position has changed. This residue is in an unusual environment, at the start of the intermolecular β-β binding region. Asp 27 N, also in both lists and Thr 30 N, in the list for the new model only, face each other in the tight loop at the start of the α helix. The other atoms in the list for the new model are Ala 53 N, which lies in the centre of the β-turn linking β strands 3 and 4 and Tyr 38 OH, which has one nearby water site.

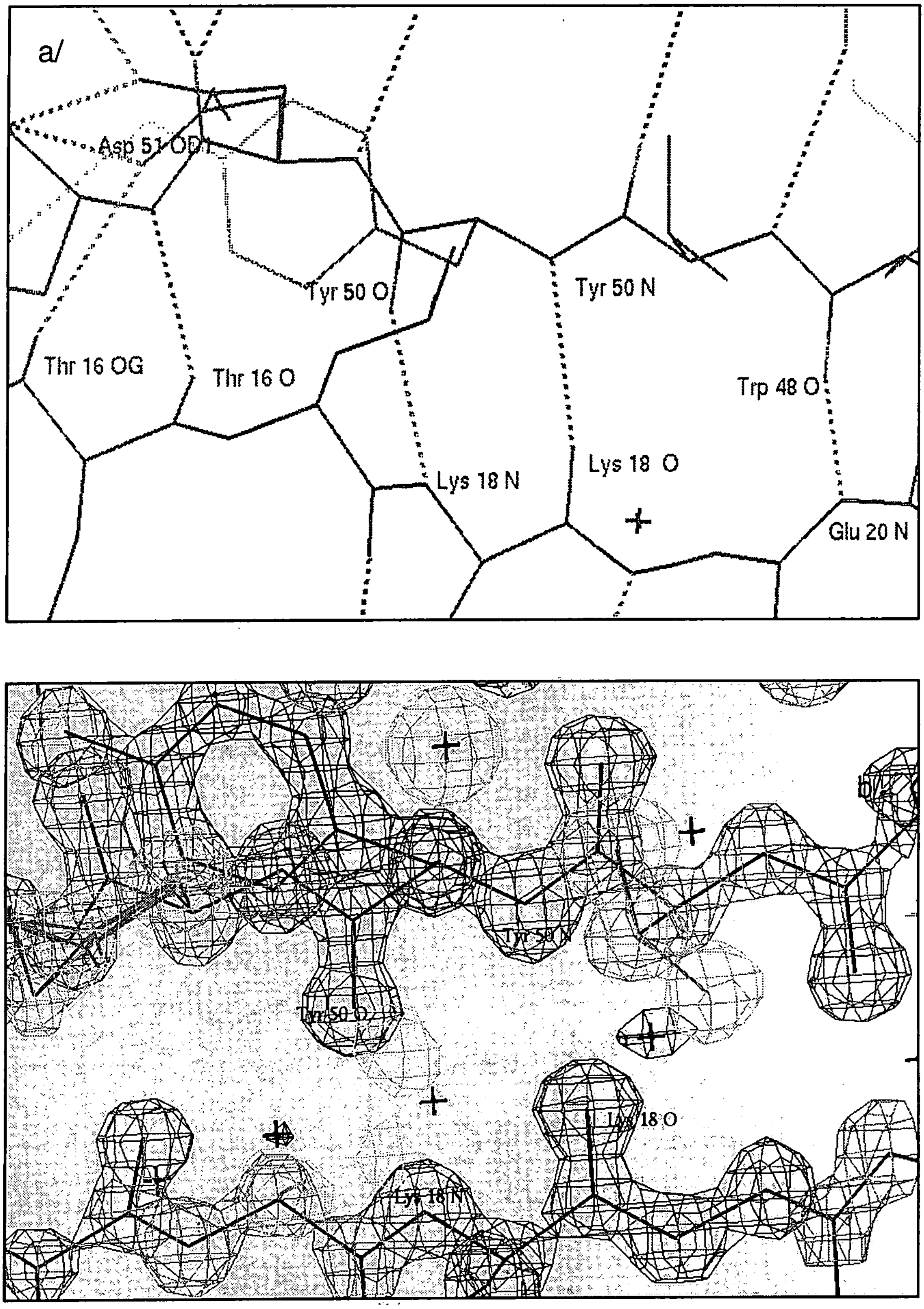


Figure 7. Intermolecular H-bonding.

a/ The five beta-beta H-bonds are shown as dashed lines.

b/ 3Fo – 2Fc electron density for this region.

Solvent

The improvement in solvent modelling can be seen from a comparison of the distributions of map correlation(1) for 1IGD and new solvent sites, Figure 8. In the new model, 50% of the solvent sites have a correlation of 0.9 - 1.0, with 18 % better than 0.95. In 1IGD, only 16 % have a correlation of over 0.9 and none greater than 0.95. Conversely, the old model contains 28 % sites with worse than 0.75 correlation, with 8 worse than 0.5, while the new model has 15 % sites with values less than 0.75, two of these below 0.5. 65 % of the solvent sites in the new model are within 1 Å of one in the 1IGD model, Figure 9. All new model sites with low B factors, < 36 Å^2, are within 0.5 Å of a site in the 1IGD model, the range of separation increasing with B factor, as illustrated by Figure 10.

From the distribution of solvent B factors, Figure 11, it can be seen that the number of sites with low B factors, 5-30 Å^2, is slightly reduced. More sites were added with B in the range 30-60 Å^2, while the number with very high B factors, over 60 Å^2, was reduced. The distribution of [(distance of solvent from protein)*B], Figure 12, shows a similar increase in sites in the intermediate range, 75-225 Å^3, while the number with extreme values declines. There are 2, possibly 3 peaks in the mid-range of the distribution, at 90 Å^3, 140 Å^3 and 220 Å^3. These peaks are more pronounced in the distribution for the new model. This suggests that during refinement the solvent in the model becomes arranged in a series of shells around the protein. The removal of peaks with low B factors close to the protein is a consequence of improved modelling of the diffuse regions of the protein and the general reduction in noise at the protein-solvent interface following the introduction of a diffuse solvent correction. The removal of peaks with very high B factors remote from the protein could be due to the reduction in noise and also to more discriminating modelling of the solvent.

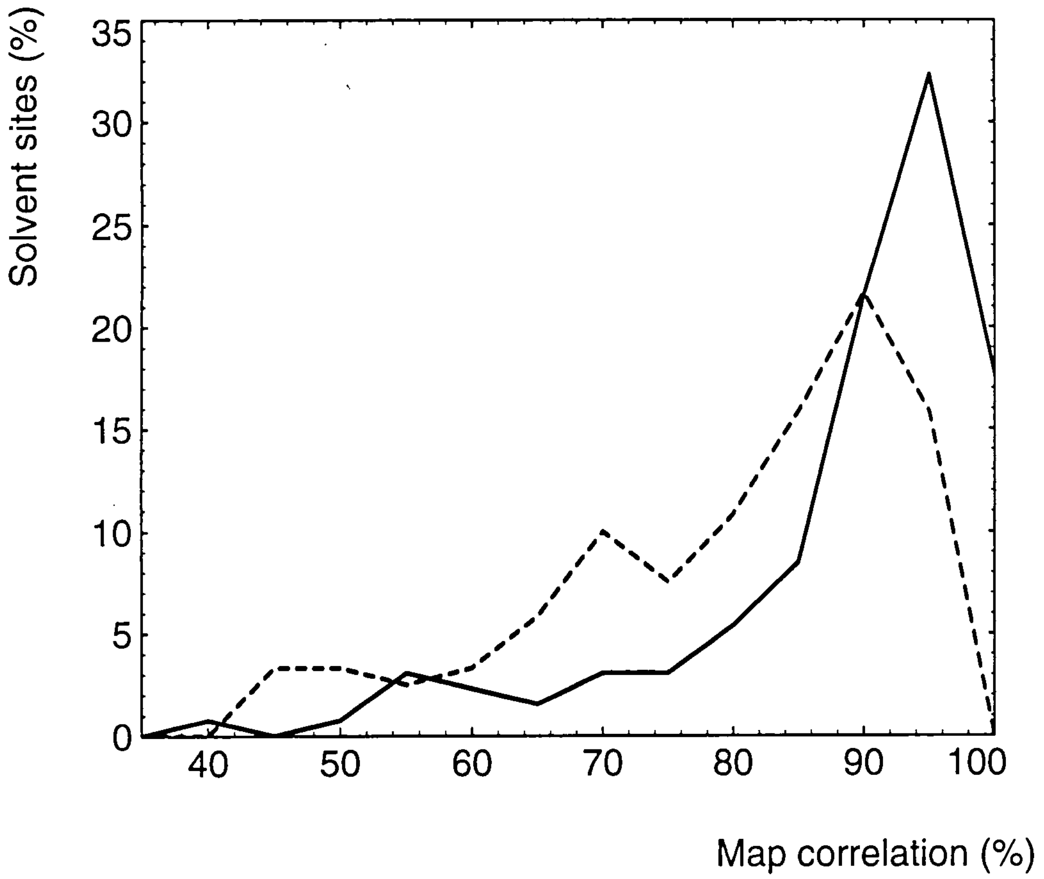


Figure 8. Map correlation of solvent sites

Histogram of correlation, k, of (3Fo-2Fc) map to Fc map for solvent sites in the anisotropic model (continuous line) and 1IGD (dashed line).

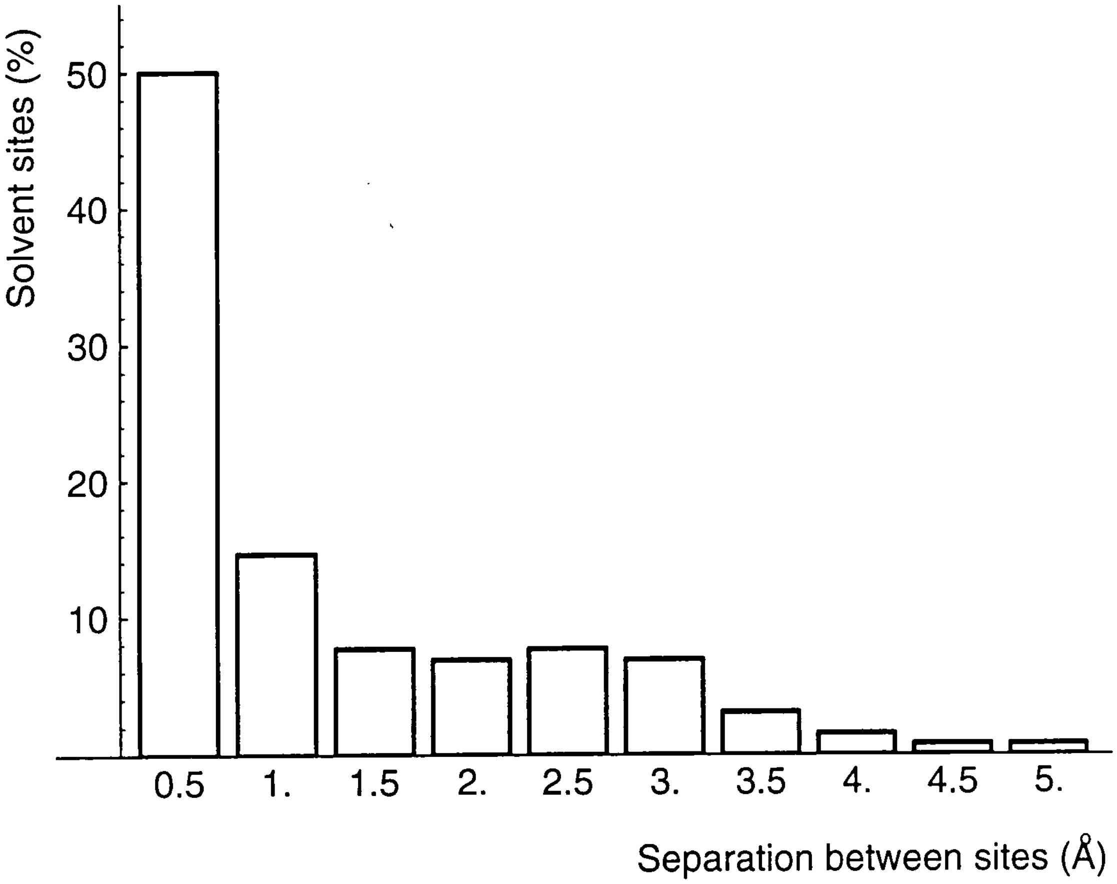


Figure 9. Separation of solvent sites

The histogram of the separation of solvent sites in the new model from those in 1IGD. The distance from each new solvent site to the closest site in 1IGD was evaluated and the distances were sorted into bins of width 0.5 Angstrom

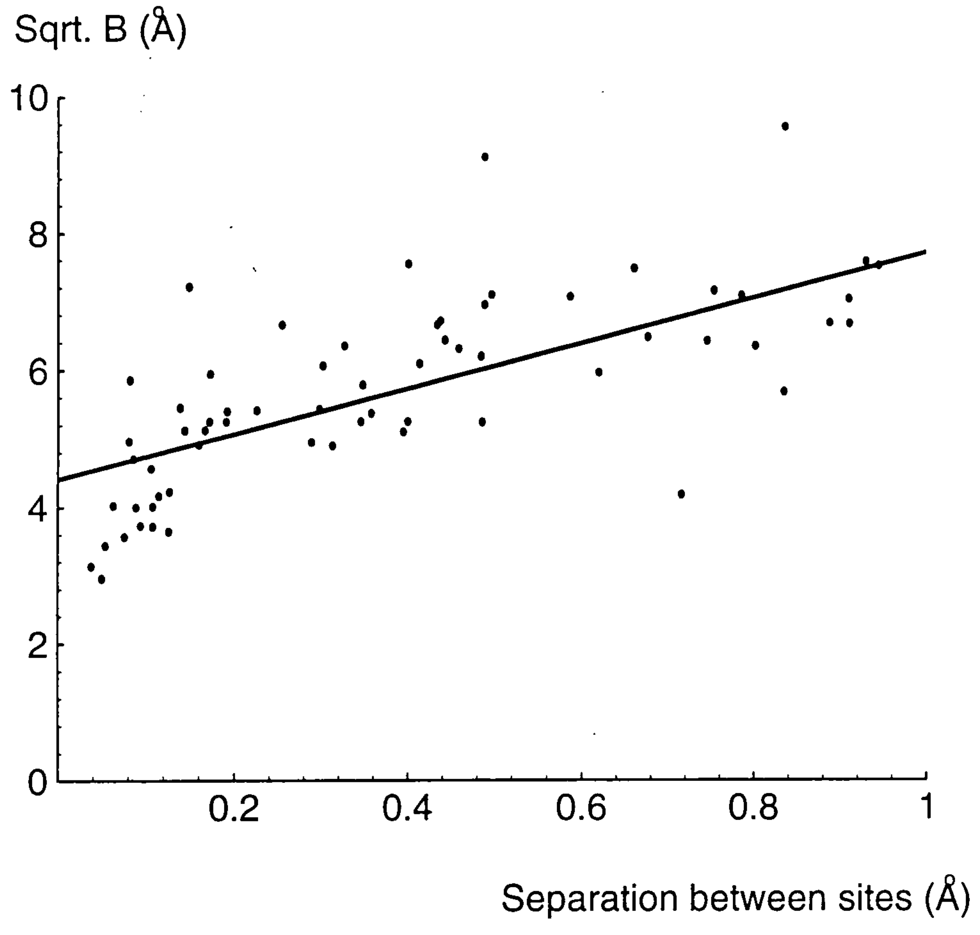


Figure 10. Separation distance between solvent sites in the 2 models against sqrt. B. The separation of each new water position from the closest 1IGD solvent site was plotted againstthe square root of the new site B factor. Only equivalent water sites, defined as those with separation less than 1 Angstrom, are shown.

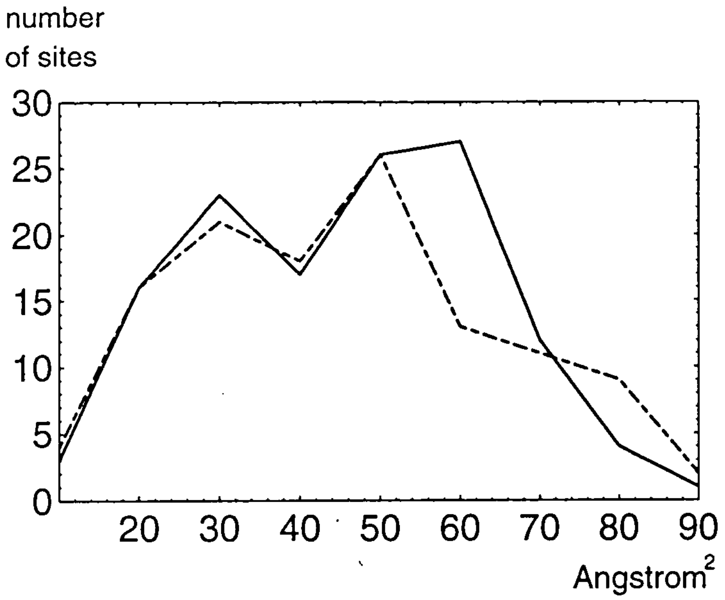


Figure 11.

New and 1IGD solvent structure are compared. Distributions of solvent site B factor are plotted for the new model (continuous line) and 1IGD (dashed line).

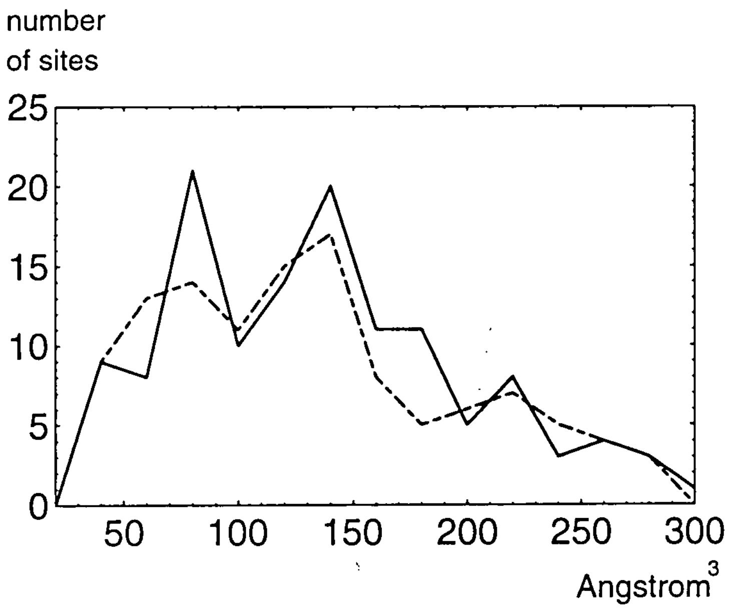


Figure 12.

Distributions of d(sp) * B(s) for the new model(continuous line) and 1IGD(dashed line) illustrate the arrangement of the solvent structure, where d(sp) is the separation of a solvent site from its closest protein atom and B(s), the solvent site B factor. The existence of a series of solvent shells, each possessing a characteristic B factor range, will result in a series of peaks in the distribution.

Immunoglobulin-binding proteins, References

Achari, A., Hale, S.P., Howard, A.J., Clore, G.M., Gronenborn, A.M., Hardman, K.D. & Whitlow, M. (1992) 1.67 Å X-ray structure of the B2 immunoglobulin-binding domain of streptococcal protein G and comparison to the NMR structure of the B1 domain. *Biochemistry* **31**, 10449-10457.

Boyle, M.D.P. (1990) *Bacterial immunoglobulin-binding proteins*, Academic press, New York.

Deisenhofer, J. (1981) Crystallographic refinement and atomic models of a human Fc fragment and its complex with fragment B of protein A from *Staphylococcus aureus* at 2.9 and 2.8 Å resolution. *Biochemistry* **20**, 2361-2370.

Derrick, J.P. & Wigley, D.B. (1994) The third IgG-binding domain from streptococcal protein G. An analysis by X-ray crystallography of the structure alone and in a complex with Fab. *J. Mol. Biol.* **243**, 906-918.

Gallagher, T., Alexander, P., Bryan, P. & Gilliland, G.L. (1994) Two structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. *Biochemistry* **33**, 4721-4729.

Gouda, H., Torigoe, H., Saito, A., Saito, M., Arata, Y. & Schimada, I. (1992) Three-dimensional solution structure of the B domain of staphylococcal protein A: comparisons of the solution and crystal structures. *Biochemistry* **31**, 9665-9672.

Gronenborn, A.M. & Clore, G.M. (1993) Identification of the contact surface of a streptococcal protein G domain complexed with a human Fc fragment. *J. Mol. Biol.* **233**, 331-335.

Gronenborn, A.M., Filpula, D.R., Essig, N.Z., Achari, A., Whitlow, M., Wingfield, P.T. & Clore, G.M. (1991) A novel highly stable fold of the immunoglobulin-binding domain of streptococcal protein G. *Science* **253**, 657-661.

Guss, B. Eliasson, M., Olsson, A., Uhlèn, M., Frej, A.-K., Jörnvall, H., Flock, J.-I. & Lindberg, M. (1986) Structure of the IgG-binding regions of streptococcal protein G. *EMBO J.* **5**, 1567-1575.

Kastern, W., Sjöbring, U. & Björck, L. (1992) Structure of peptostreptococcal protein L and identification of a repeated immunoglobulin light chain-binding domain. *J. Biol. Chem.* **267**, 12820-12825.

Kihlberg, B.-M., Sjöbring, U., Kastern, W. & Björck, L. (1992) Protein LG: a hybrid molecule with unique immunoglobulin-binding properties. *J. Biol. Chem.* **267**, 25583-25588.

Lian, L.-Y., Yang, J.C., Derrick, J.P., Sutcliffe, M.J., Murphy, J.P., Goward, C.R. & Atkinson, T. (1991) Sequential ^{1}H assignments and secondary structure of an IgG-binding domain from protein G. *Biochemistry* **30**, 5335-5340.

Lian, L.-Y., Derrick, J.P., Sutcliffe, M.J., Yang, J.C. & Roberts, G.C.K. (1992) Determination of the solution structures of domains II and III of protein G from *Streptococcus* by ^{1}H NMR. *J. Mol. Biol.* **228**, 1219-1234.

Lian, L.-Y., Barsukov, I.L., Derrick, J.P & Roberts, G.C.K. (1994) Mapping the interactions between streptococcal protein G and the Fab fragment of IgG in solution. *Nature Struct. Biol.* **1**, 355-357.

Stryer, L. (1988) *Biochemistry, 3rd edition.* W.H. Freeman & Co. New York.

Torigoe, H., Shimada, I., Saito, A., Saito, M. & Arata, Y. (1990) Sequential 1H NMR assignments and secondary structure of the B protein of staphylococcal protein A: structural changes between the free B domain in solution and the Fc-bound domain in crystal. *Biochemistry* **29**, 8787-8793.

see also general crystallographic references, end of Chapter 5

Chapter 4B:

Anisotropic refinement of rubredoxin from *Desulfovibrio vulgaris*

Background

Rubredoxin is a small bacterial non-haem iron-containing protein, around 50 amino acids in length, Figure 1. The iron atom is complexed to four cysteine sulphur atoms in an approximately tetrahedral geometry. The protein has been isolated from sulphate-reducing and some aerobic bacteria, however the aerobic rubredoxin is a larger molecule of around 80 residues. Its precise function is not known, but it is thought to participate in a redox chain. Electron transfer interactions with cytochrome c_3 have been reported (Bell *et al.*, 1978). Extensive structural studies have been carried out on rubredoxins from several species. Table 1 lists crystal structures which have been determined.

Table 1. Rubredoxin crystal structures

source	*best resol. (Å)*	*best R (%)*	*residues*	*H_2O*	*symmetry/cell*	*references*
Clostridium pasteurianum ox, Fe	1.1	9.0	54	110	R3: 64.0, 64.0, 32.5 Å	Watenpaugh *et al.*, 1973 Watenpaugh *et al.*, 1979 Dauter *et al.*, 1996
Clostridium pasteurianum ox, Zn	1.2	10.7	54	87	R3 64.1, 64.1, 33.1 Å	Dauter *et al.*, 1996
Desulfovibrio vulgaris	1.0 (1.5)	14.7 (9.8)	52	102	$P2_1$ 19.9, 41.4, 24.4 Å 108.3°	Pierrot *et al.*, 1976 Adman *et al.*, 1991 Dauter *et al.*, 1992 Sheldrick *et al.*, 1993
Desulfovibrio gigas	1.4	14.0	53	117	$P2_1$ 19.7, 41.7, 24.4 Å 109.4°	Pierrot *et al.*, 1976 Frey *et al.*, 1987
Desulfovibrio desulfuricans	1.5	9.3	45	121	P1 24.9, 17.8, 19.7 Å 101, 83.3, 104.5°	Sieker *et al.*, 1986 Stenkamp *et al.*, 1990
Pyrococcus furiousus ox	1.8	17.8	53	61	$P2_12_12_1$ 34.6, 35.5, 44.4 Å	Day *et al.*, 1992
Pyrococcus furiousus red	1.8	19.3	53	37	$P2_12_12_1$ 34.6, 35.5,.44.4 Å	Day *et al.*, 1992

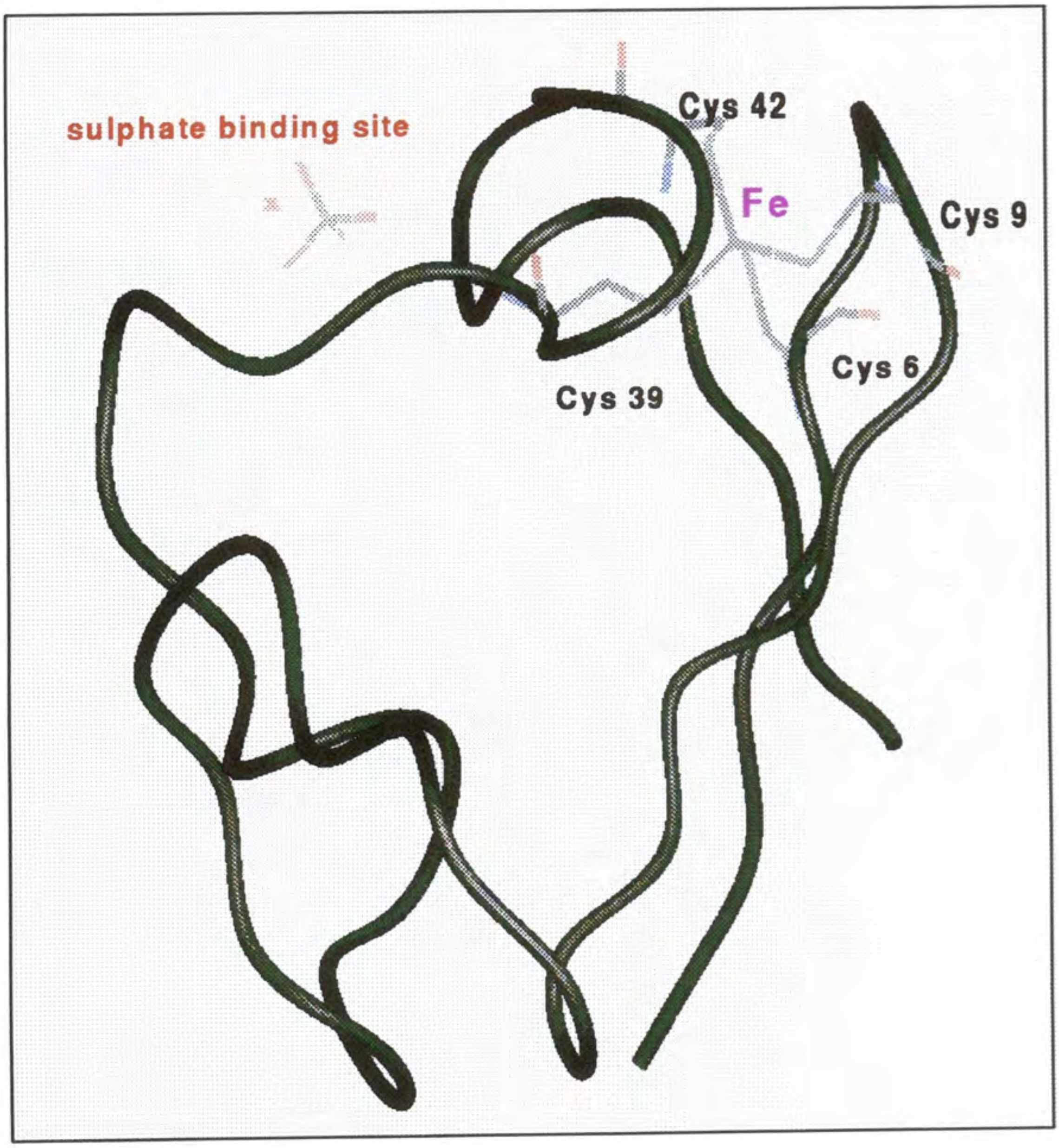


Figure 1. Overall fold of rubredoxin , showing the Fe (Cys)4 cluster and the sulphate binding site. Depicted using QUANTA.

The redox potential of rubredoxin varies from -60 to +6 mV between species (Moura *et al.*, 1979). Structural studies have attempted to correlate changes in sequence and activity and explain how structure is conserved with changes in sequence. Rubredoxin forms crystals which are often of exceptional quality. This has resulted in the protein becoming the subject of much work aiming towards the development of methods in protein crystallography, including pioneering work on refinement of protein structures by least-squares minimisation (Watenpaugh *et al.,* 1973) and the application of small molecule structure solution methods to protein data (Sheldrick *et al.*, 1993).

Aims

Rubredoxin from *Desulfovibrio vulgaris* is a member of a set of proteins, for all of which X-ray diffraction data had been collected at EMBL Hamburg, which were being used as test structures during the ongoing development of refinement and evaluation techniques in protein crystallography. The refinement of the structure relied heavily on the automated methods under development. A comparison was made between the anisotropic model at 0.92 Å and the completely independent, 1.0 Å resolution, isotropic model published by Dauter *et al.*, (1992). This allowed the assessment of advantages to be gained by the adoption of new refinement protocols and the identification of areas for future development of techniques.

Experimental

The refinement of Rubredoxin was undertaken as part of this PhD. The crystallisation, data collection and refinement to give the starting model in the present study were performed previously.

Crystallisation and Data Collection

Dark red prismatic crystals were grown as described by Adman *et al.*, (1977) from a 0.5 - 1.0% protein solution, buffered by 0.1 M sodium citrate at pH 4, with 2 M ammonium sulphate precipitant. Crystal properties are listed in Table 2. Data were collected as outlined by Sheldrick *et al.*, (1993) on the X31 beamline, with a wavelength of 0.70 Å. The detector was a MAR research imaging plate scanner. Data processing statistics are summarised in Table 2.

Table 2. Crystal properties and data processing statistics

Crystal symmetry;	*Monoclinic, $P2_1$*
Molecules per asymmetric unit	1

Cell dimensions (Å)	a= 19.99, b=41.51, c=24.40
	β = 107.60 °

V_m (Å^3 da^{-1})	1.74
Resolution (Å)	20 - 0.92
R_{merge} (%)	3.7
N_{uniqve} (Freidels separate)	48 291
N_{unique} (Freidels merged)	26 108
Completeness (%)	98.9

Initial Model and Data

The starting model for this refinement was derived from an unpublished, intermediate model obtained by George Sheldrick. This model had been refined anisotropically using SHELXL-93 (Sheldrick, 1993) to an R factor of 7.44 %. The model comprised 52 residues, one Fe atom, a sulphate ion and 117 water sites. The sulphate ion, with an associated water, had 68% occupancy, while another water, with 32 % occupancy, also lies at the sulphur atom position. The remaining water sites numbered 63 fully occupied, 13 pairs with 50:50 occupancy and 26 single sites with half occupancy, making a total of 90 water molecules in the asymmetric unit. 6 residues: Lys 2, Glu 12, Pro 15, Asp 21, Lys 25 and Ser 29 had two alternate side chain conformations, with all hydrogen atoms included.

To obtain the starting model for the new refinement, the sulphate ion and all water molecules were removed, as well as the minor conformations of disordered residues, the major conformation occupancy being reset to 100 %. Co-ordinates of the remaining atoms were randomised by rms 0.3 Å to reduce model bias from reflections now to be used for the evaluation of R_{free}. The data consisted of a total of 48291 unique reflections, 26108 with Freidel pairs merged, in the resolution range 20-0.92 Å. Freidel pairs were held separately in the reflection file, but merged before refinement and map calculation. 5% of the reflections, 1327, were removed from the working dataset and used for the calculation of R_{free}, the refinement being independent of these data. The commencing stages of the refinement were against Fs, while in following stages intensities were used.

Refinement strategy

Restrained least-squares refinement of atomic positions and thermal parameters was performed. The initial stages consisted of isotropic refinement, against Fs, using PROLSQ. When the isotropic model had converged satisfactorily, anisotropic refinement against intensities using SHELXL-93 followed. Stereochemical restraints using ideal values taken from the Engh & Huber (1991) set were applied during both the PROLSQ and SHELXL-93 stages of the refinement, as explained in the introduction to this chapter. Hydrogen atom positions were not refined, but calculated using a riding model. The Automated Refinement Procedure (ARP, Lamzin & Wilson, 1993) was used continuously during the refinement for construction and improvement of the solvent network.

The scheme of refinement is summarised in Table 3. Initially, ten cycles of PROLSQ were run on the randomised model, partly as a further precaution to remove any remaining memory of the R_{free} data. The density for the sulphate ion was then clearly apparent in the ($3F_o$-$2F_c$) and difference maps, so it was built into the model to be refined without restraints. Following five more cycles of PROLSQ, the construction of the solvent network commenced. A cycle of ARP was run after each PROLSQ cycle, with ARP set to modify solvent only, adding and removing up to ten atoms per cycle, within a range of 2.2-3.3 Å from existing atoms. Solvent atoms which came closer than 0.5 Å to one another were merged. The refinement converged with R 14.9 %, R_{free} 18.3 % and 99 water sites.

Table 3. Scheme of refinement

Cycle		R (%)	ΔR	R_{free} (%)	ΔR_{free}	water sites	$\rho_{rms, F}$ #	$\rho_{rms, \Delta F}$ ##
-	randomisation of co-ordinates from previous refinement	33.7	-	34.6	-	0		
1-10	isotropic refinement, PROLSQ	21.1	*-12.6*	21.8	*-12.8*	0	0.823	0.163
11-15	introduction of sulphate ion	20.3	*-0.8*	21.7	*-0.1*	0	0.818	0.157
16-30	PROLSQ & construction of solvent network with ARP	14.9	*-5.4*	18.3	*-3.4*	99	0.745	0.097
	isotropic refinement , SHELXL-93	15.9	*+1.0*			99		
31-40	anisotropic refinement , SHELXL-93, ARP	8.5	*-7.4*	11.5	*-6.8*	91	0.695	0.076
41-50	6 double conformation side chains and 2 conformations for 45-48, with occupancy 70:30 %	8.1	*-0.4*	10.7	*-0.8*	83	0.682	0.072
51-60	sulphate/glycol modelled in the sulphate site, sulphate and major split chain conformation refined together	7.9	*-0.2*	11.0	*+0.3*	85.7*	0.683	0.071
61-70	refinement against 100 % data	7.9	*0*	(8.4)**	*(-2.6)***	85.7*	0.699	0.071

$\rho_{rms, F}$ = rms of ($3F_o$-$2F_c$) map density * the 0.7 is a water bound to the sulphate

$\rho_{rms, \Delta F}$ = rms of (F_o - F_c) map density ** these data are no longer independent

Before anisotropic refinement, five cycles of SHELXL-93 were run on the isotropic model. The R factor rose to 15.9 %, due to the application of a different set of restraints, as explained in the introduction to this chapter. This rise in R factor is a commonly observed phenomenon. Anisotropic refinement proceeded with ten rounds, each consisting of five cycles of SHELX-93 refinement followed by a cycle of ARP. The R factor dropped to 8.5% and R_{free} to 11.5%. Following these ten rounds of anisotropic refinement, the fit of the model into the ($3F_o$-$2F_c$) and difference maps was examined using FRODO (Jones, 1978). Difference density showed the positions for second conformations of 6 residues: Met 1, Glu 12 , Pro 15, Asp 21, Lys 25 and Ser 29. Two alternative conformations for each side chain were modelled. H atoms bound to partially occupied C, N and O atoms were not included in the model. Geometric restraints on the residues with multiple conformations were tightened. The occupancies of the alternate conformations were refined, with the sum of occupancies for each residue set to 100 %. Following further refinement, a third position for Ser 29 OG became apparent. Three conformations were then modelled, with the occupancies set to 75%, 20% and 5%.

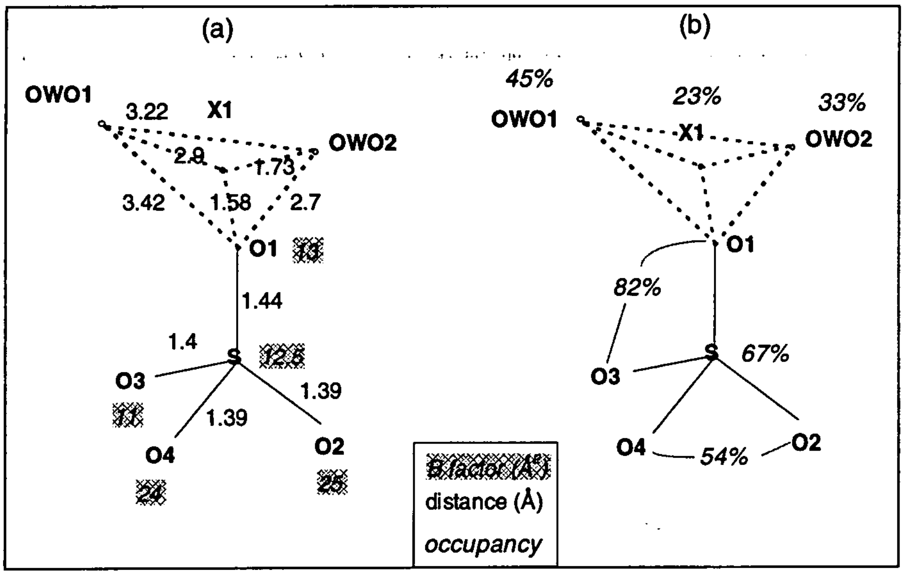
At this stage, most of the difference density was clustered in one region of the map, the cavity containing the sulphate ion, and the surrounding section of chain, residues 45-48. Inspection of the difference density around these residues clearly indicated the position of a second main chain conformation. The largest difference peak, X1, was situated 1.58 Å from O1 of the sulphate. The sulphate itself, which had been refined from the beginning with no restraints, was evidently imperfectly modelled. Although the geometry was tetrahedral, the bond lengths were abnormally short for sulphate and the density for O2 and O4 was very weak, see Figure 2a. The search for an appropriate way to treat this problem comprised several steps.

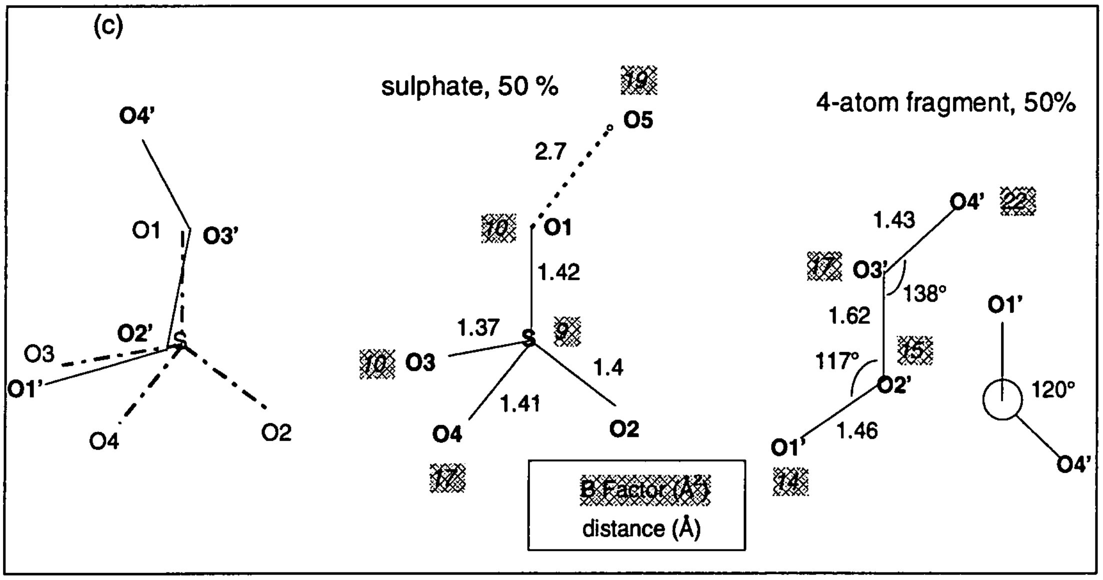
First, alternative main chain and side chain conformations for residues Pro 45, Lys 46, Ser 47 and Glu 48 were modelled. Restraints were set allowing independent movement of the two alternate chain conformations, with distance restraints between the two conformations completely removed, since stronger restraints were found to pull the minor conformation back on to the major. The occupancies of the two parts were fixed at 70% and 30%. These changes were introduced to the model following the first ten rounds of anisotropic refinement, after cycle 40. Ten rounds of refinement were run on the model with the alternate chain conformation, then a further ten, with the occupancies of the alternate chain conformations and the sulphate now refined, but independently of each other. The major chain occupancy refined to 73% and the sulphate to 66%. Ten further rounds of refinement followed, in which the occupancy of an oxygen at X1 was also refined. The occupancies of chain, sulphate and X1 were then 75%, 62% and 25%.

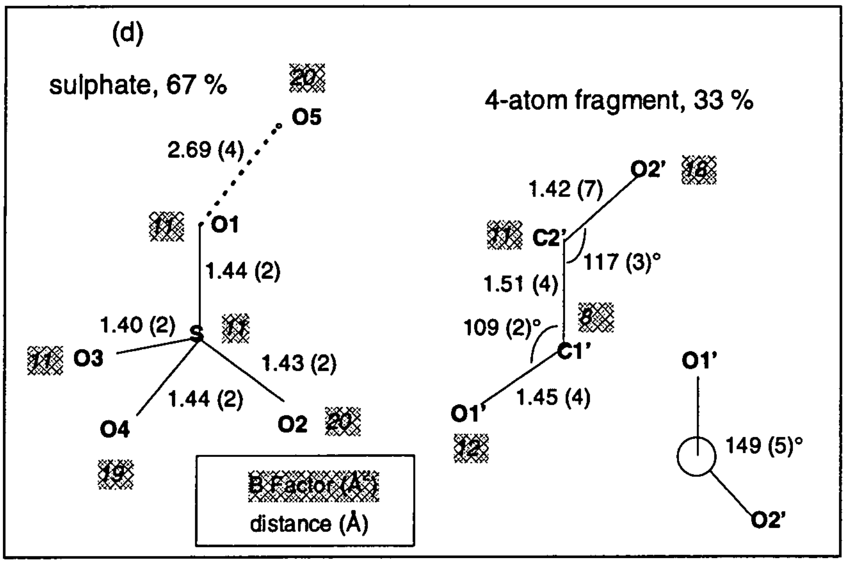
From cycle 50, the occupancies were refined separately for the chain, X1, the sulphate S, sulphate O1 and O3, sulphate O2 and O4 and two nearby water sites OWO 1 and OWO 2. The resulting occupancies are shown in Figure 2b. Next, four atoms, modelled as oxygen, were placed in the sulphate site, in the absence of sulphate. OWO 2 is clearly bound to the sulphate O1, so it was assigned the same occupancy as the sulphate and renamed O5. The relative occupancies of the four-atom fragment and sulphate were then refined. The result of this step is shown in Figure 2c.

The geometry of the four-atom fragment was consistent with that of glycol, although the crystal had not been exposed to glycol at any stage. It was suggested that the four atoms visible were part of a citrate ion, the remainder of which was completely disordered. The four atoms were modelled as glycol, with appropriate restraints applied. The sulphate bond lengths were restrained to be equal and the geometry, tetrahedral. The occupancies of the split chain and sulphate/glycol were refined independently to 70% / 30%. Finally, the occupancies of the split chain and sulphate/glycol were refined as one variable, from cycle 50. After ten rounds of refinement, the sulphate and major chain conformation occupancy was 67(1)%. The final sulphate and glycol conformations are shown in Figure 2d.

Figure 2, stages in modelling of the disordered sulphate ion







The ambiguity has not been completely removed, but much more information has become apparent. In the final model, the mean separation of pairs of main-chain atoms in the split section were 0.46 Å, 0.59 Å, 0.53 Å and 0.31 Å for residues 45, 46, 47, and 48 respectively. The sulphate ion is present for 67 % of the time. When the sulphate is absent, the protein chain around the site shifts by around 0.5 Å to adopt an alternative conformation, thus complexing more effectively with the species now occupying the cavity, with Ser 47 OG, N46 and N47 positioned to interact and also Lys 46 NZ, although the second conformation of the lysine side chain is not certain.

The largest difference peaks which remained, following the alterations described above, were all in the solvent region. This reveals the inadequacy in the solvent model, consisting of 85 fully occupied water sites. If the relative volumes of the protein and the unit cell it occupies are considered, it can be estimated that there is sufficient space for approximately 94 waters. Observation of the density in the solvent region reveals 'dumb bell' and 'chain-of-beads' type features, with distances between neighbouring peaks in the 1.0 - 2.4 Å range. The solvent structure evidently consists of alternate networks of partially occupied sites. Therefore, although the number of waters modelled is appropriate, a model with more sites, some with partial occupancies, would be better. Preliminary attempts were made to assign half occupancies to pairs of sites, with difference peaks greater than 4σ, separated by less than H-bonding distance, but no change in R or R_{free} was observed and the atoms tended to drift out of the density during subsequent refinement. More stringent modelling criteria are evidently required.

The working and R_{free} datasets were combined and ten rounds of refinement were run against all the reflections. This precaution was taken because random absences in a dataset cause ripples in density maps and such spurious density may result in incorrect modelling of the structure. In this case, refinement with the extra 5% of data made no significant difference. Errors in the final model were estimated by running a cycle of overlapping blocked refinement, as explained in the introduction. As for protein G, these errors had a strong correlation with the atomic B factors. Mean co-ordinate errors estimated were 0.0152 Å, 0.0121 Å and 0.0129 Å for main-chain atoms, CA, N and O.

Comparison with an isotropic model of the rubredoxin structure

The 1.0 Å resolution structure of rubredoxin from *Desulfovibrio vulgaris* (Dauter *et al*, 1992), PDB code 8RXN, obtained by an independent refinement of a different data set, provides an instructive comparison to the new model. The resolution range of the present dataset is greater, with more high and low resolution reflections included. However, the principle difference is that the present structure is modelled with anisotropic thermal parameters for all non-hydrogen atoms, while the 8RXN structure is an isotropic model. Differences between the models are summarised in Table 4.

Table 4. Comparison between 1.0 Å isotropic (8RXN) and 0.92 Å anisotropic models

	8RXN		New Refinement
Cell dimensions (Å)	a= 19.97, b=41.45, c=24.41 β = 108.3 °		a= 19.99, b=41.51, c=24.20 β = 107.6 °
N_{unique}	18 532		26 108
Completeness (%)	90.5		98.9
R_{merge} (%)	5.8		3.7
Resolution (Å)	12.8 - 1.0		20.0 - 0.92
R (%)	14.7		7.9
total water sites	102		85.7
B protein, mean (Å^2)	10.0		9.1
B residue, mean (Å^2)	10.7		9.6
B SO_4, mean (Å^2)	16.3		14.5
B Fe (Å^2)	5.1		4.9
B H_2O, mean (Å^2)	38.8		32.1
rmsd, CA (Å)		0.058	
rmsd, protein (Å)		0.149	
mean H_2O sep. (Å)		0.523	

Unit Cell

The 8RXN model was refined using a cell derived from the indexing of synchrotron data. These measurements possess an inherent uncertainty resulting from the uncertainty in the measurement of the X-ray wavelength. For data measured at Hamburg in the era of the collection of the rubredoxin and protein G data sets which this chapter is concerned with (pre mid-1995) there may be an error of up to 0.5% in cell dimensions. Since this time, more stringent procedures have been introduced, entailing regular calibration of the X-ray wavelength and crystal to detector distance by measurement of silicon powder diffraction rings. This has reduced the inherent error in cell dimensions from synchrotron measurements to around 0.1 %, which is sufficiently small to have no significant effect on the accuracy of bond distances in a structure. The present structure was refined using the cell obtained from diffractometer measurements (Adman *et al.*, 1977). This is not an infallible practice, since there are examples of significant variation in crystal cell dimensions, even between crystals grown in a single drop.

A systematic deviation in the 8RXN model bond lengths was noted. Overestimation of cell dimensions was one suggested explanation, it could also be related to the use of a different (pre Engh & Huber) dictionary for the application for geometric restraints. Exactly the same phenomenon was observed in the comparison of isotropic and anisotropic refinements of protein G (Chapter 4a).

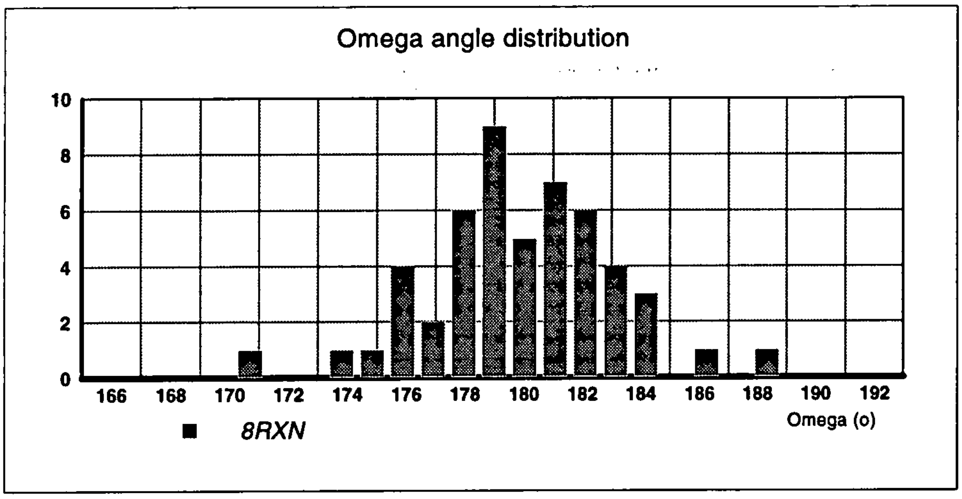
Protein

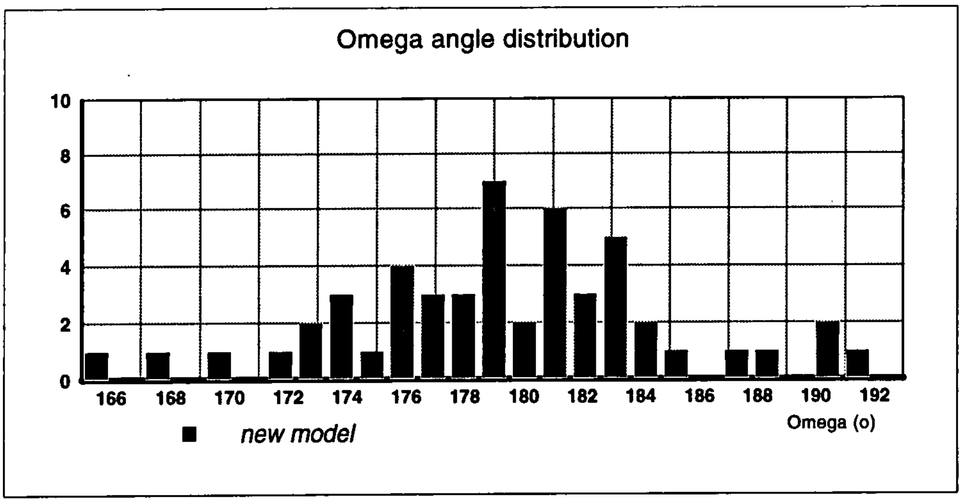
A comparison of the overall geometry of new and 8RXN models, Table 5, shows that average bond distances in the 8RXN model are longer, as mentioned above. The width of the distributions of bond lengths was also greater for 8RXN, indicating that different restraints were applied during the refinement. The distribution of the peptide bond torsion angle, Ω, is much more sharply peaked for 8RXN and centred at 180°, while the new model has a much wider range of Ω values and a slightly lower mean value, Figure 3. The deviation from planarity for peptide bonds is a phenomenon observed in high resolution protein structures, in the absence of over-strong planarity restraints.

Table 5. A comparison of overall geometry of new and 8RXN models

	New model		target (Å)	8RXN	
	mean interatomic distance (Å)	σ interatomic distance (Å)		mean interatomic distance (Å)	σ interatomic distance (Å)
N-CA	1.455	0.014	1.458	1.464	0.027
CA-C	1.516	0.016	1.525	1.520	0.035
C-O	1.228	0.012	1.231	1.231	0.034
C-N	1.324	0.013	1.329	1.325	0.030
N-C	2.458	0.034	2.462	2.458	0.039
CA-O	2.387	0.018	2.401	2.391	0.034
O-N	2.233	0.016	2.250	2.244	0.021
C-CA	2.429	0.015	2.435	2.429	0.030
CA-N	2.427	0.022	2.425	2.421	0.027
	mean Ω (°)	σ[Ω] (°)	target (°)	mean Ω (°)	σ[Ω] (°)
	179.7	5.1	180.0	179.9	3.1

Figure 3





The rmsd between the models is 0.149 Å for all protein atoms and 0.058 Å for CA atoms. The rmsd was calculated with the nomenclature of atoms adjusted to be identical for the two structures, with changes to Tyr 4, Asp 21, Phe 30 and Asp 31. The largest rmsd's are seen for residues in regions of diffuse density. Met 1, Glu 12, Glu 17, Asp 21, Asp 31, Glu 50 and Ala 52 all have rmsd greater than 0.2 Å, Lys 3 and Pro 15, greater than 0.1 Å, Figure 4. The B factors of protein atoms in the new model are, on average, 1 Å^2 smaller, while much larger differences exist for residues in regions of diffuse density, Figure 4*a*. The residues with the largest reductions in B factor between old and new models, in descending order, are Lys 2, Ala 52, Glu 17, Glu 12, Asp 31, Met 1, Asp 21 and Pro 15.

The N-terminus in the new model has an additional formate group. There is density in the ($3F_o$-$2F_c$) map for this group, however there is some residual negative density in the (F_o-F_c) map. During the 8RXN refinement, broad, low atomic peaks were seen, indicating disorder, but this was not modelled. Two conformations for the side chain have been modelled in this refinement, the major conformation being very similar to that in the old model. The B factors are around 2 Å smaller. Lys 2 and Lys 3 were also noted to be disordered in the previous study and CD, CE and NZ of Lys 2 were assigned zero occupancies. All atoms were included in the new model, however the density for the Lys 2 side chain remains extremely weak and there is negative difference density at the Lys 3 side chain position.

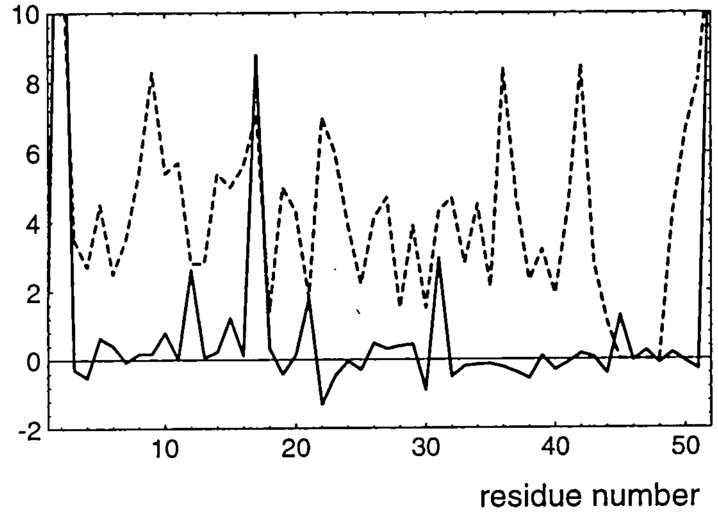
The side chains of Glu 12, Pro 15 and Asp 21 were all modelled with two alternate conformations in both cases. In the 1 Å structure, the occupancies were set to 50:50, while for the new model, the major conformation occupancy refined to 60(2)%, 67(3)%

and 58(2)%. The conformations of residues 15 and 21 are very similar, with the refinement of occupancy and anisotropic modelling of atoms resulting in reduction in B factors. Neither of the Glu 12 side chain conformations is conserved between the models. The new conformations appear to fit into the density better and the atoms have B factors on average 3 Å^2 lower.

During the 8RXN refinement, peaks in the difference synthesis suggested the existence of more than one conformation for several other residues, but the density was not sufficiently well defined to allow their modelling. Some of this difference density could result from the uncompensated-for anisotropy of the atoms and possibly some from the insufficient treatment of diffuse solvent, see below. The new refinement includes two conformations for Lys 25 and three for Ser 29 which were not previously modelled, although the second conformation of Ser 29 was present as a water site.

Figure 4. Comparison of the final model with 8RXN.

(a) The differences between the 2 models are plotted against residue number; [B(residue,old)-B(residue,new)] (continuous line), 100[rmsd CA atoms] (dashes)



(b) The rms distances between equivalent atoms in the 2 models are plotted against residue number; 100[rmsd(CA)] (continuous line), 100[rmsd(all atoms)] (dashes)

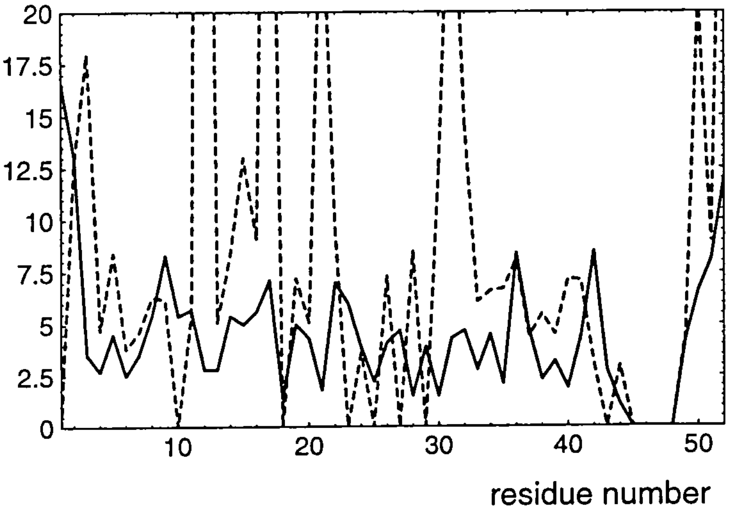
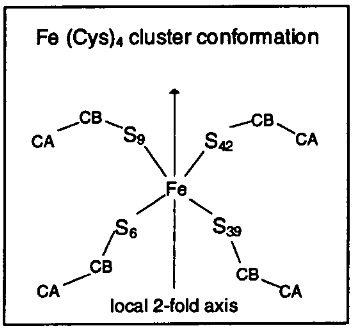


Table 6. Comparison of [Fe(Cys)$_4$] cluster geometry

bond lengths (Å)	8RXN	new model	bond angles (°)	8RXN	new model
Fe-S_6	2.30	2.295 (3)	S_6-Fe-S_9	115	114.5 (1)
Fe-S_{39}	2.30	2.295 (3)	S_{39}-Fe-S_{42}	112	111.7 (1)
Fe-S_9	2.27	2.258 (2)	S_6-Fe-S_{39}	111	110.8 (1)
Fe-S_{42}	2.27	2.254 (2)	S_9-Fe-S_{42}	109	110.1 (1)
			S_6-Fe-S_{42}	106	105.5 (1)
			S_9-Fe-S_{39}	105	104.4 (1)
torsion angles (°)					
Fe-S_6-CB-CA	180	-173.1 (3)	Fe-S_6-CB	101	102.2 (2)
Fe-S_{39}-CB-CA	180	-175.3 (4)	Fe-S_{39}-CB	99	100.3 (2)
Fe-S_9-CB-CA	270	-88.9 (5)	Fe-S_9-CB	110	109.8 (2)
Fe-S_{42}-CB-CA	270	-94.7 (5)	Fe-S_{42}-CB	110	108.7 (2)

bond lengths (Å)	8RXN	new model		8RXN	new model
Fe-N_8	5.13	5.15 (1)	S_6-N_8	3.54	3.56 (1)
Fe-N_9	3.77	3.80 (1)	S_6-N_9	3.55	3.57 (1)
Fe-N_{11}	4.69	4.77 (1)	S_9-N_{11}	3.42	3.47 (1)
Fe-N_{41}	5.10	5.10 (1)	S_{39}-N_{41}	3.57	3.57 (1)
Fe-N_{42}	3.87	3.89 (1)	S_{39}-N_{42}	3.62	3.64 (1)
Fe-N_{44}	4.84	4.83 (1)	S_{42}-N_{44}	3.50	3.49 (1)
bond angles (°)					
S_6-N_8-CA	98.8	98.6 (3)	CB-S_6-N_8	106.1	106.5 (2)
S_6-N_9-CA	122.0	121.9 (3)	CB-S_6-N_9	101.7	100.5 (2)
S_9-N_{11}-CA	111.1	109.5 (4)	CB-S_9-N_{11}	113.8	114.8 (2)
S_{39}-N_{41}-CA	104.7	103.8 (4)	CB-S_{39}-N_{41}	106.4	106.3 (2)
S_{39}-N_{42}-CA	120.7	118.2 (4)	CB-S_{39}-N_{42}	95.4	95.0 (2)
S_{42}-N_{44}-CA	105.6	105.5 (4)	CB-S_{42}-N_{44}	114.2	115.5 (3)



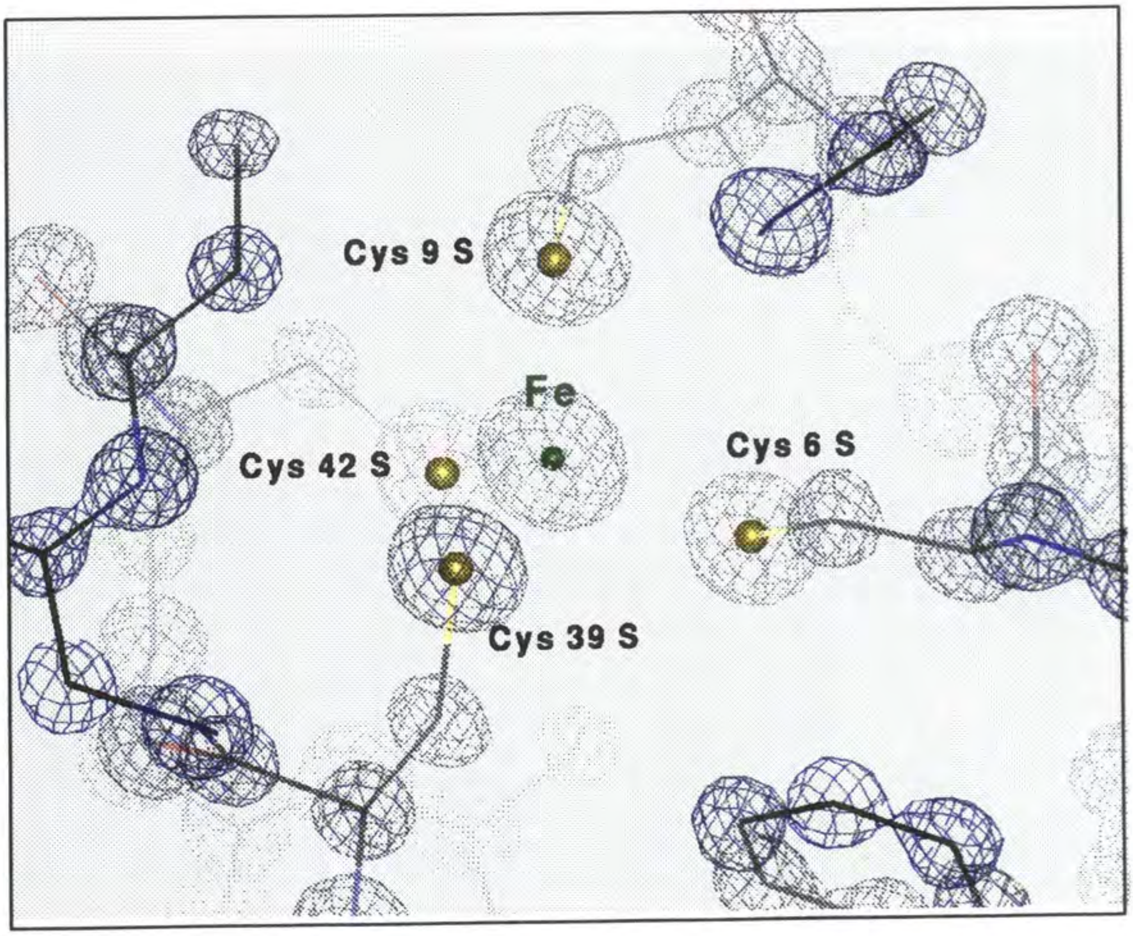


Figure 5. The Fe(Cys)4 cluster. 3Fo – 2Fc density.

Plotted using QUANTA

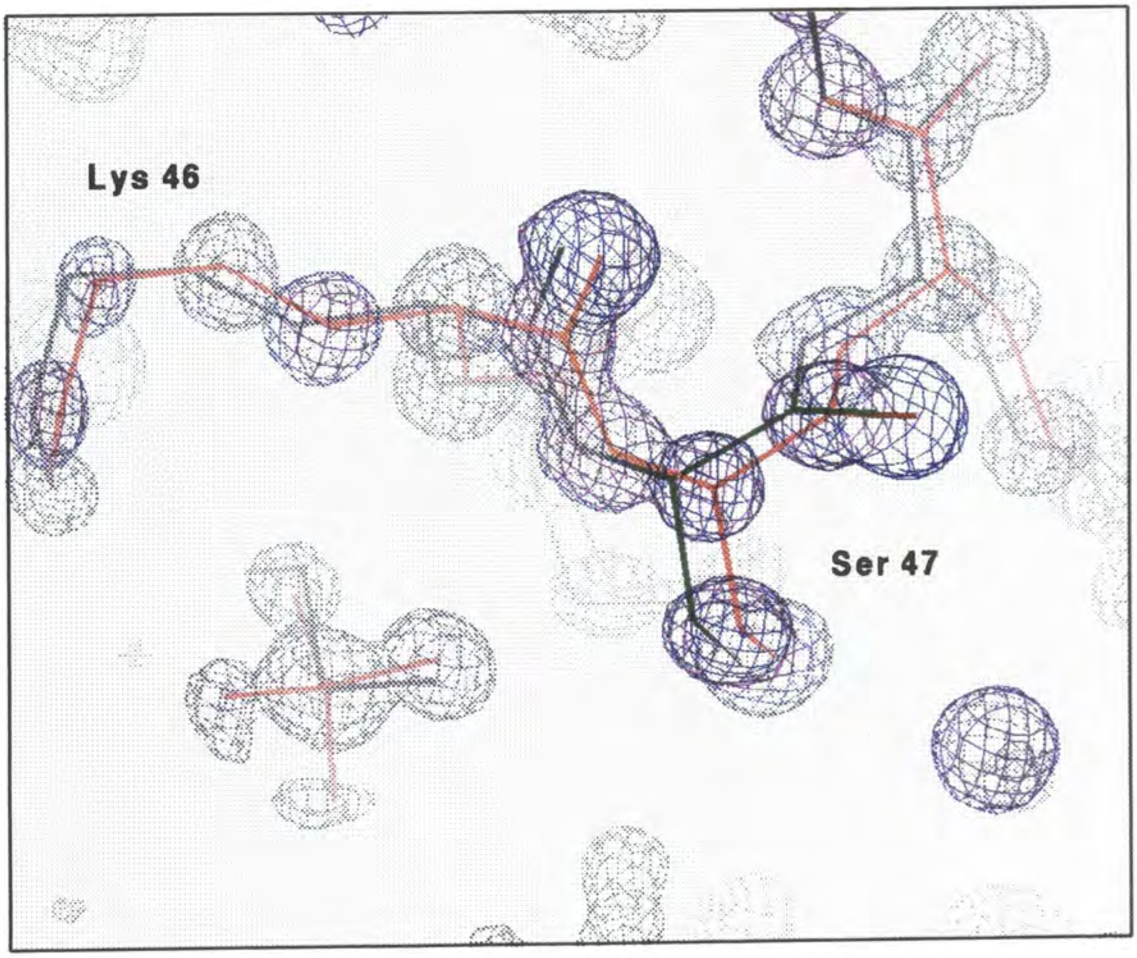


Figure 6. The sulphate binding site. The major chain conformation and sulphate are shown in red, the minor chain conformation, in green. The (3Fo - 2Fc) electron density is plotted. Using QUANTA.

Glu 17 and Asp 31 have diffuse side chain density in old and new structures, however anisotropic modelling has improved the fit to the density for both residues. Residues Asp 32 and 33, as well as the C-terminal Ala 52, are also shifted from their positions in the old model, resulting in a better fit to the density.

The [Fe(Cys)$_4$] cluster, Figure 5, adopts close to perfect C_2 symmetry, with pairs of sulphur atoms: S_6 & S_{39}, S_9 & S_{42} related by the local 2-fold axis. The symmetry extends to include two sets of three nitrogen atoms; N8, N9, N11 and N41, N42 & N44, which surround the cluster, forming N-H···S bonds. A comparison of the two models, Table 6, shows no significant differences and the rmsd of FeS_4 between the two structures is 0.03 Å.

The major chain conformation for residues 45-48 in the new model was similar to that of the 1 Å model. The B factors per residue are also similar. However, the old and new position and geometry of the sulphate ion, Figure 6, are different, Table 7.

Table 7. Comparison of $SO_4{}^{2-}$ geometry

bond lengths (Å)	8RXN	new model	bond angles (°)	8RXN	new model
S-O$_1$ (S-O$_3$) *	1.55	1.44 (2)	O$_1$-S-O$_3$	113	111 (1)
S-O$_3$ (S-O$_1$) *	1.46	1.40 (2)	O$_2$-S-O$_4$	111	107 (2)
S-O$_2$	1.51	1.43 (2)	O$_1$-S-O$_2$ (O$_3$-S-O$_2$) *	109	107 (1)
S-O$_4$	1.54	1.44 (2)	O$_3$-S-O$_4$ (O$_1$-S-O$_4$) *	109	111 (1)
B factors (Å^2)			O$_1$-S-O$_4$ (O$_3$-S-O$_4$) *	104	105 (1)
S	12	11	O$_2$-S-O$_3$ (O$_2$-S-O$_1$) *	108	114 (1)
O$_1$ (O$_3$) *	13	11			
O$_3$ (O$_1$) *	11	11			
O$_2$	22	19	* O$_3$ (8RXN) ≡ O$_1$ (new)		
O$_4$	23	20			

The rms deviations between old and new models are 0.087 Å, 0.098 Å, 0.137 Å, 0.173 Å, 0.284 Å, for S, O1, O3, O2 & O4 respectively. In the old model, the S-O bond lengths averaged 1.52 Å, in the new, only 1.43 Å, a lower than expected value for sulphate. In the new model, the S-O1 bond was the longest, prior to application of restraints and modelling of a partially occupied glycol molecule, while in the old model it was the bond equivalent to S-O3. A puzzling difference in B factors between (O1, O3, S) and (O2, O4) exists in both models and the peak seen in the difference map at 1.6 Å from O1, prior to the modelling of glycol in the new refinement, Figure 2a, is also observed in the other model. On balance, the evidence suggests that different treatment of the sulphate during the two refinements, rather than differences in the environment of the molecule in the crystal, has given rise to the differences in the models.

The structure validation program WHATCHECK (Vriend & Sander, 1993) searches for potential hydrogen bonding atoms buried inside the protein, which are not participating in hydrogen bonds. Virtually all protein atoms in the structure were classed as buried, due to the close packing of the crystal. The list of unsatisfied hydrogen bond donors and acceptors was the same for both models, with the exception of Lys 2 NZ, which is not present in 8RXN. The density around the Lys 2 side chain remains too diffuse for the modelling to be completed. The other atoms listed are Lys 46 N and Lys 25 NZ, which are both disordered and pointing, from opposite directions, into the sulphate binding site.

Solvent

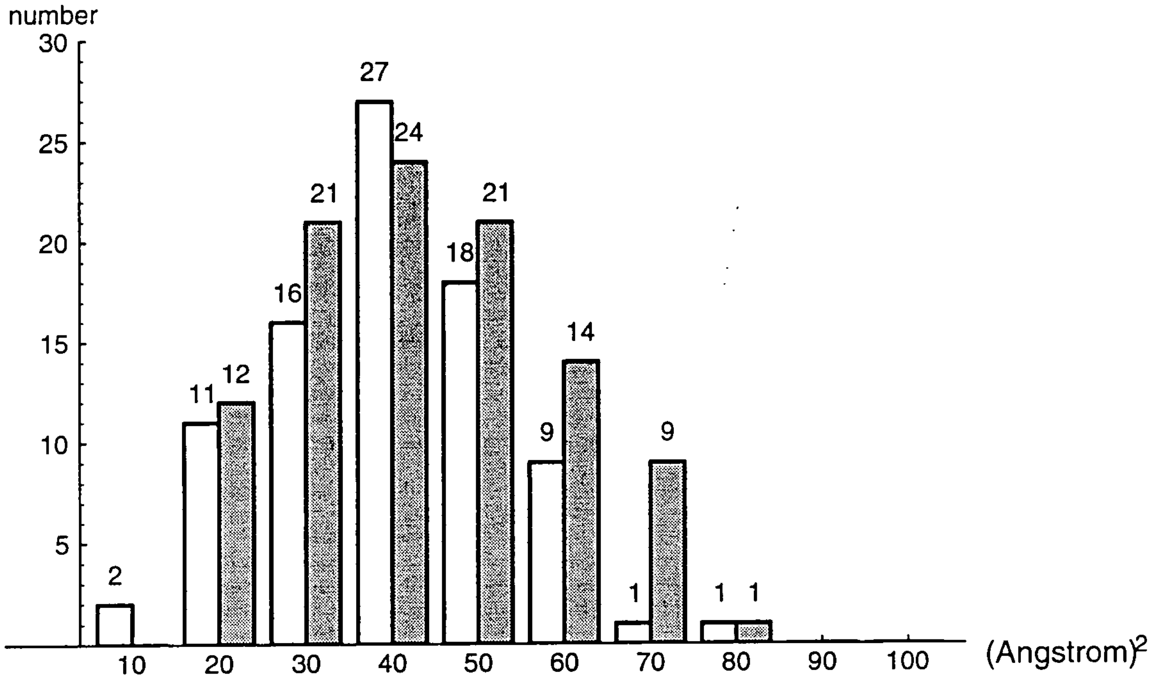
The 8RXN model contains 102 water sites, compared to 85 in the new model, all fully occupied. The mean separation of a site in the new model from one in the old is 0.52 Å and over 90 % of new sites are within 1.0 Å of an old site, Figure 7*a*. Solvent B factors are, on average, 4 Å^2 lower in the new model. Dauter *et al.* (1992) noted a surprising deficiency of water sites with low B factors, only 12 with B < 20 Å^2. The new model contains 13 such sites. The distribution of solvent B factors, Figure 7b, is more sharply peaked for the new model, with 32 % of sites with mid-range B factors, 30-40 Å^2, compared to 23.5 %. Overall, the number of sites with medium and low B factors is similar for both models, while the number with high values, > 40 Å^2, is larger in the old model, which has 45 such sites compared to 29.

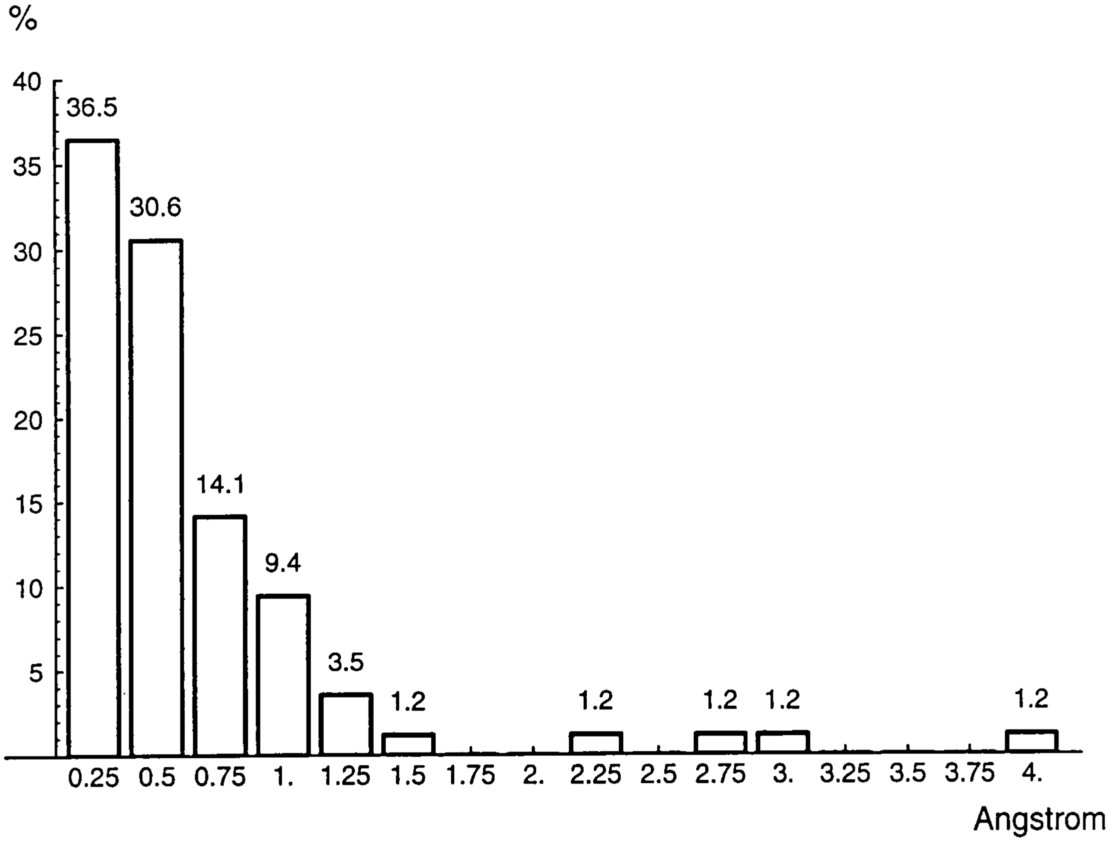
By considering the relative molecular and asymmetric unit volumes, it can be estimated that there is space for approximately 94 water molecules in the asymmetric unit. There are actually more than 94 water sites, since many are partially occupied. The 8RXN model contains more of these disordered water sites, with full occupancies assigned. An attempt was made during that refinement to model solvent with partial occupancies, assigned on the basis of their B factors, but this achieved no reduction in R factor, so it was abandoned. In this refinement, the disordered solvent network was not modelled and this resulted in most of the remaining difference density peaks being located in the solvent region, as described above.

During this refinement, density maps were calculated using data which had been scaled to account for the contribution of diffuse solvent to the low resolution scattering, below about 4 Å. This effect is not as large for rubredoxin as for most protein structures, since the tight crystal packing leaves little room for a diffuse solvent region. This is reflected by the V_m for the structure, 1.74 Å^{3}Da^{-1} (Matthews,1968). For protein crystals V_m values lie in the range 1.65 to 3.35 Å^{3}Da^{-1}, most frequently between 2.1 and 2.4 Å^{3}Da^{-1}. Thus, rubredoxin lies at the lower limit of the distribution. The effect of application of a diffuse solvent correction to data prior to map construction is to reduce the noise in the density at the protein-solvent interface. This is one factor contributing to the reduction in number of solvent sites found during the new refinement compared to the old.

Figure 7. Comparison of the solvent structure of the final model with 8RXN

(a) distributions of solvent B factors for the new model (white) and 8RXN (grey).





(b) The distance of each solvent site in the new model from the closest solvent site in 8RXN was evaluated. The distribution of these separations is plotted, with frequencies given as percentages of the total solvent in the new model.

General

WHATCHECK tabulates pairs of atoms separated by an unusually short distance. 38 of these 'bumps' were listed for 8RXN, 25 between water sites, 13 between solvent and protein atoms. The new model had 6 water/water bumps, 6 water/protein and 1 protein/protein. The large number of bumps for 8RXN results from the large number of water sites, with 100% occupancy assigned, added to weak density. The bumps in the new model were inspected using FRODO. The closely separated water sites included a chain of three and three pairs of sites which would be better modelled with partial occupancies. The solvent sites approaching protein too closely were in regions of diffuse density around the Lys 3, Glu 12 and Glu 17 side chains. The pair of protein atoms listed is Pro 26 C and Lys 25 O, separated by 2.77 Å. This appears to be a genuinely short distance, brought about by a tight bend in the chain centred around Pro 26.

Anisotropic refinement made improvements in the modelling possible. During the 8RXN refinement, residual peaks caused by anisotropic thermal motion were observed in the difference map. This was most clearly illustrated by the doughnut-shaped features lying round the atoms of the [Fe(Cys)$_4$] cluster. The modelling of this anisotropy in the present refinement greatly reduced the noise in the difference map, increasing the visibility of features in weak density regions. The effect of this was seen in the improved positioning of several weakly scattering residues such as Asp 31, Asp 32 and Asp 33 and, most notably, by the appearance of difference peaks on the main chain for residues 45-48, which indicated the position of a minor chain conformation. The modelling of atoms anisotropically, coupled with the ability to refine occupancy, allowed multiple conformations to be modelled more effectively, thus more multiple conformations could be built into the model, for example, two conformations for Met 1, and they refined to give a better fit to the density, as can be seen by a comparison of old and new models for Asp 21.

Other questions which need to be addressed include whether the increased resolution range of the data in the new refinement has lead to any obvious improvement. The contribution of the extra data is not obvious, since the 13 - 1 Å range was sufficient for a high resolution refinement. The effect of introducing anisotropy to the model is much larger and probably masks any changes caused by the presence of extra data. However, if the 1 Å model were to be modelled anisotropically, the effect of the additional data would probably become more noticeable and the R factor would be higher for the 1 Å than the 0.92 Å resolution structure.

Another question is whether it is possible to distinguish between divergence in the models arising from the distinct refinement protocols and differences actually present in the crystal structures. This was mentioned in relation to the modelling of the sulphate site. Genuine differences are small and therefore masked by the much larger effects resulting from the specific treatment of the models. To be certain of this point, it would be necessary to refine the two structures with an identical approach, but in isolation from one another, so the features observed for one model do not give clues about how to model the other.

Rubredoxin References

Adman, E.T., Sieker, L.C., Jensen, L.H., Bruschi, M. & Le Gall, J.A. (1977) Structural model of rubredoxin from *D. vulgaris* at 2 Å resolution. *J. Mol. Biol.* **112**, 113-120.

Adman, E.T., Sieker, L.C. & Jensen, L.H. (1991) Structure of rubredoxin from *Desulfovibrio vulgaris* at 1.5 Å resolution. *J. Mol. Biol.* **217**, 337-352.

Bell, G.R., Lee, J-P., Peck, H.D., Jr. & Le Gall, J.A. (1978) Reactivity of *Desulfovibrio gigas* hydrogenase toward artificial and natural electron donor or acceptors. *Biochimie* **60**, 315-329.

Dauter, Z., Sieker, L.C. & Wilson, K.S. (1992) Refinement of rubredoxin from *Desulfovibrio vulgaris* at 1.0 Å with and without restraints. *Acta Crystallogr.* **B48**, 42-59.

Dauter, Z., Wilson, K.S., Sieker, L.C., Moulis, J.M. & Meyer, J. (1996) Zn and Fe-rubredoxins from *Clostridium pasterianum* at atomic resolution: the first high precision model of ZnS_4 unit in a protein. *Proc. Natl. Acad. Sci. USA, in press.*

Day, M.W., Hsu, B.T, Joshua-Tor, J.B., Park, Z.H., Zhou, M.W., Adams, M.W. & Rees, D.C. (1992) X-ray crystal structures of the oxidised and reduced forms of the rubredoxin from the marine hyperthermophilic archaebacterium *Pyrococcus furiosus. Protein Science* **1**, 1494-507.

Frey, M., Sieker L.C., Payan, F., Haser, R., Bruschi, M., Pepe, G. & Le Gall, J.A. (1987) Rubredoxin from *Desulfovibrio gigas.* A molecular model of the oxidised form at 1.4 Å resolution. *J. Mol. Biol.* **197**, 525-541.

Moura, I., Moura, J.J.G., Santos, M.H., Xavier, A.V. & Le Gall, J.A. (1979) Redox studies on rubredoxins from sulphate and sulphur reducing bacteria. *FEBS Lett.* **107**, 419-421.

Pierrot, M., Haser, R., Frey, M., Bruschi, M., Le Gall, J.A., Sieker, L.C. & Jensen, L.H. (1976) Some comparisons between two crystallised anaerobic bacterial rubredoxins from *Desulfovibrio gigas* and *D. vulgaris. J. Mol. Biol.* **107**, 179-182.

Sheldrick, G.M., Dauter, Z., Wilson, K.S., Hope, H. & Sieker, L.C. (1993) The application of direct methods and Patterson interpretation to high-resolution protein data. *Acta Crystallogr.* **D49**, 18-23.

Sieker, L.C., Stenkamp, R.E., Jensen, L.H., Prickril, B.C. & Le Gall, J.A. (1986) Structure of rubredoxin from the bacterium *Desulfovibrio desulfuricans. FEBS Lett.* **208**, 73-76.

Stenkamp, R.E., Sieker, L.C. & Jensen, L.H. (1990) The structure of rubredoxin from *Desulfovibrio desulfuricans* strain 27774 at 1.5 Å resolution. *Proteins: Struct. Funct. Genet.* **8**, 352-364.

Watenpaugh, K.D., Sieker, L.C., Herriot, J.R. & Jensen, L.H. (1973) Refinement of the model of a protein. Rubredoxin at 1.5 Å resolution. *Acta Crystallogr.* **B29**, 943-956.

Watenpaugh, K.D., Sieker, L.C. & Jensen, L.H. (1979) The structure of rubredoxin at 1.2 Å resolution. *J. Mol. Biol.* **131**, 509-522.

Chapter 5:

Aspects of the refinement of atomic resolution protein structures

A: Sharpening as a tool in protein crystal structure refinement

Summary

Due to the fall-off of scattering intensity with increasing θ, high resolution data are generally weak. Structure factors can be normalised to remove their resolution dependence, so the high resolution terms make a more significant contribution to electron density maps, resulting in sharper atomic peaks. Since the high resolution intensities are inherently weak measurements, they have relatively large associated errors, thus their upweighting leads to a concurrent magnification of the associated errors and the appearance of spurious peaks in the density map. An optimal degree of sharpening of the data leads to maps in which atomic peaks are sharp and well defined, while the noise contribution is minimal. The desirable degree of sharpening varies with characteristics of the structure including the resolution of the data, overall B factor and associated errors.

This study involved an estimation of the most informative level of sharpening for maps of varying resolution and quality through an evaluation of the properties of the electron density distributions. Refinement runs were performed to test these deductions. It was concluded that a degree of sharpening applied to maps during refinement was indeed beneficial at all resolutions at which atoms can be distinguished in the density. The optimal amount of sharpening was found to vary between 100% for the preliminary stages of a refinement at atomic resolution to around 50% in the later stages.

Introduction

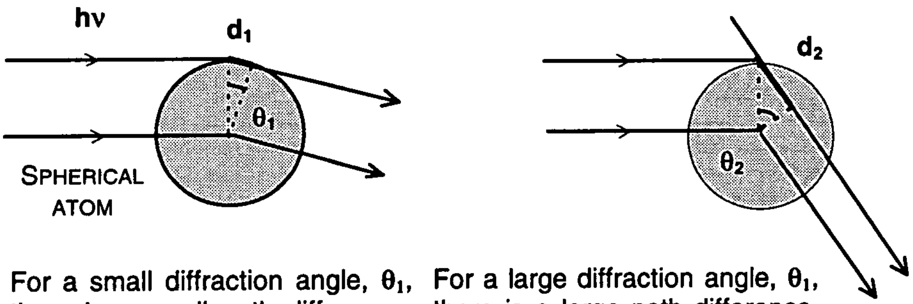
Structure Factors, F's & E's

A structure factor is the result of the diffraction, in a specific direction which satisfies the Bragg conditions, from all parts of the electron density within the unit cell of a crystal structure. For the calculation of model structure factors this is approximated by the sum of the scattering from all the atoms within the unit cell, as expressed in equation (1). The structure factor possesses phase and amplitude. For calculated structure factors, the phase relates to the atomic position relative to the cell origin and the amplitude to the nature of the electron distribution around the atom.

$$F(\mathbf{h}) = \Sigma_{j=1,N} f_j \, exp[\, 2\pi i(\mathbf{h}x_j)] \qquad (1)$$

The atomic scattering factor is resolution dependent. Due to the finite size of the electron charge cloud around an atom, the path difference between waves diffracted from different parts increases with the diffraction angle, resulting in destructive interference, Figure 1.

Figure 1.
Schematic illustration explaining the cause of fall-off in diffraction intensity with increasing resolution for spherical atoms of a finite size.



For a small diffraction angle, θ_1, there is a small path difference, $\mathbf{d_1}$ between diffracted waves

For a large diffraction angle, θ_1, there is a large path difference, $\mathbf{d_2}$ between diffracted waves

The scattering factor of an atom in a real crystal (2) declines even more rapidly with increasing scattering angle, due to the spreading of the atomic electron density by static and dynamic effects. The scattering factor for a point atom (3) is without resolution dependent terms.

$$f_j = f_{o,j} \exp[-Bs^2] \qquad (2)$$

where $s = \sin\theta/\lambda$ and $B = 8\pi^2u^2$, u^2 = mean squared amplitude of atomic vibration.

at the limit of (2), when $\theta = 0$, $f_j = f_{o,j}(\theta = 0) = Z_j$, Z = atomic number

$$f_{\text{point atom}} = Z_j \qquad (3)$$

For a cell containing equal atoms, (1) becomes $F = f\,\Sigma X$, with $X = exp[\,2\pi i(\mathbf{hx})]$ and the relation between the structure factors for a real and a point atom can be obtained by combining (2) and (3) to give (4). For structures containing j different atoms, the approximation (5) can then be made. The normalised structure factor, E(**h**) obtained from F(**h**), (6) (Karle & Hauptmann, 1956), has all the resolution dependent terms cancelled out, so can be thought of as the scattering factor for a point atom.

$$F_{\text{point}} / F_{\text{real}} = Z / f_o \exp[-Bs^2] \qquad (4)$$

$$F_{\text{point}} = F_{\text{real}} \ (\Sigma Z_j) / (\ exp[-Bs^2]\ (\Sigma f_{o,j})\ ) \qquad (5)$$

$$E(\mathbf{h}) = F(\mathbf{h}) / (< F^2 >)^{1/2} \qquad (6)$$

A normalisation factor may be determined, assuming randomly distributed atoms, from the gradient of the Wilson plot (Wilson, 1942). The ideal intensity of a reflection is given by (7), where f_i and f_j are atomic scattering factors. If reflections are divided into small ranges of resolution and the mean intensity is calculated, the exponential terms will tend to zero, leaving a simple relation (8). The ideal intensity differs from the real intensity by two factors, a scaling factor independent of resolution, k, and a thermal factor. Thus, the ratio of ideal to real intensity is a function of these two factors. With the resolution dependent temperature factor expressed as exp $[-2Bs^2]$ this leads to (9). The plot $\ln[<I_{hkl}>/\Sigma_i f_i^2\,]$ versus s^2 has gradient -2B and intercept ln k.

$$I_{hkl} = \Sigma_i \Sigma_j f_i f_j \, exp[\, 2\pi i\, (\, h(x_i - x_j) + k(y_i - y_j) + l(z_i - z_j)\,)\,] \tag{7}$$

rearranged $I_{hkl} = \Sigma_i f_i^2 + \Sigma_{i \neq j} \Sigma f_i f_j \, exp[\, 2\pi i\, (\, h(x_i - x_j) + k(y_i - y_j) + l(z_i - z_j)\,)\,]$, ...

$$<I_{hkl}> = \Sigma_i f_i^2 \tag{8}$$

$$\ln\, [<I_{hkl}> / \Sigma_i f_i^2\,] = \ln k - 2Bs^2 \tag{9}$$

This derivation assumes that the structure is composed of a collection of uniformly distributed atoms whose thermal motion may be described by an overall B factor. The first assumption is very approximate for a protein crystal (Blessing & Langs, 1988, 1996). At low resolution, the structure is divided into higher density protein regions, interspersed with lower density solvent regions, with solvent occupying between 25% and 65% of the volume (Matthews, 1968). At medium resolution the regular secondary structural features, typically adopted by around 75% of the molecule, are observed. Bonding and non-bonding distances fall into a narrow range. The result is that the Wilson plot will possess a series of characteristic fluctuations. Realistic estimates of B and k may be obtained from the plot for data excluding the lower resolution range at which these fluctuations are particularly significant.

The description of thermal motion by an overall B factor is improved if anisotropy is accounted for using a matrix, **Uij**. Use of the mean B factor for scaling would be appropriate if the B factor distribution for a structure were Gaussian, the mean and mode values for a Gaussian distribution both being equivalent. In fact the distribution is skewed, with a sharp cutoff for low values and a long tail for high values. As a result, the mean B value is larger than the peak value and it may be better to use the latter (Blessing & Langs, 1996).

The K-curve method of estimating E values (Karle & Hauptmann, 1953; Blessing & Langs, 1988) entails the division of the data into resolution bins. A scale factor, K is calculated for each bin (10). The values of K are plotted against s and smoothing is applied. The function K = *f*(s) so obtained gives the scaling factors. This method accounts for the deviation of the structure from a uniform distribution of atoms.

$$K(s) = < \Sigma f_j^2 (s) / | F|^2 >_s \qquad (10)$$

$$E_{hkl} = K\, F_{hkl} \qquad (11)$$

Electron Density Maps

The electron density map in real space and the diffraction data in reciprocal space are related by a Fourier transform (12). To generate an electron density map, both magnitude and phase of the structure factors are required. Phases are calculated by the inverse Fourier transform from the atomic model. Observed and/or calculated amplitudes are input. During a crystal structure refinement, fitting of calculated to observed data in reciprocal space is typically alternated with fitting of the model structure into electron density maps in real space, by either automated analysis of the density or manual inspection using computer graphics. The quality of these maps is therefore crucial to the success of the refinement.

$$\rho(x,y,z) = (1/V)\, \Sigma_h\, \Sigma_k\, \Sigma_l\, |F_{hkl}|\, exp[-2\pi i(hx + ky + lz - \alpha'_{hkl})] \qquad (12)$$

where $\alpha' = \alpha/2\pi$, the phase angle in cycles

The contribution of high resolution data to the F_o synthesis is decreased by the reduction in magnitude of structure factors as the diffraction angle increases. E values are not resolution dependent, so E maps are more strongly influenced by the high resolution terms. Comparison of E and F maps shows that the former have sharper, better defined atomic peaks. Thus E-values are 'sharpened' structure factors. Since weak reflections have large errors associated with their measurement, upweighting of low intensity high angle data also magnifies errors, resulting in a higher noise level in the sharpened map. However combining E and F, giving 'semi-sharpened' structure factors, allows the advantages of sharpening to be exploited while the drawbacks are minimised.

Experimental

Models

This study involved four crystal structures, Table 1. The data were collected using synchrotron radiation at EMBL. The models were refined using similar protocols with details described elsewhere. The resolution of the data covers the range over which the sharpening of maps might be expected to be beneficial, from atomic to around 2 Å.

Table 1. Crystal Structures

Structure	*Spacegroup*	*Resolution (Å)*	*Wilson Plot B factor (Å^2)*	*Reference*
Rubredoxin	$P2_1$	20.0 - 0.92	15	Dauter *et al.*, 1992
Protein G	$P2_12_12_1$	10.0 - 1.10	20	Derrick & Wigley, 1994
Eglin c	$P4_3$	10.0 - 2.00	37	Betzel *et al.*, 1993
Transthyretin	$P2_12_12$	10.0 - 1.90	44	Damas *et al.*, 1996

Maps

$(X_o\text{-}X_c,\alpha_c)$ and $(3X_o\text{-}2X_c,\alpha_c)$ maps, where $X = (F^x E^{(1-x)})$ and $x \leq 1$, with varying sharpness and resolution limits, were computed, using FFT (Ten Eyck, 1973), ECALC and other programs from the CCP4 (1994) suite. Following the application of resolution cuts to the data, 20 cycles of restrained least-squares refinement were run using the CCP4 version of PROLSQ (Konnert & Hendrickson, 1980) to help remove the memory of the high resolution data from the model.

Both observed and calculated structure factors were then independently normalised to give E-values. Calculation of E-values necessitates the division of the data into a sufficient number of bins with an adequate number of reflections in each bin, so care is required in setting bin widths for data sets to which successively large high resolution cutoffs have been applied. This is particularly important for structures with a small asymmetric unit, such as rubredoxin, protein G and eglin c, due to the relatively small number of low resolution reflections. Prior to the calculation of semi-sharpened maps, a scale factor of $(\Sigma F^2/\Sigma E^2)^{1/2}$ was applied to the E-values, so that maps calculated with different degrees of sharpness would be comparable. Such scaling gives rise to a comparable rms density for maps of varying sharpness at a given resolution.

A further problem with the scaling of low resolution data, especially those originating from a synchrotron, can arise due to strong reflections overloading the detector and thus being wrongly measured. The only solution is to be observant during data collection and to collect a high resolution data set in three or four runs, as explained in Chapter 4B.

Nominal and Effective Resolution of the Data

An electron density map of a protein crystal structure consists of a collection of blurred peaks, corresponding to the electron density around atoms. The amount of detail which can be visualised in a density map is determined by the resolution of the data. However, the nominal resolution is not the only deciding factor. The peaks are broadened by thermal motion and static disorder. Thus, the quality of a map generated from data artificially cut to a certain resolution will be different to that of one from data collected from crystals which only diffracted to that resolution.

The first ideas about atom resolvability in electron density maps obtained from X-ray diffraction were formulated for the theoretical case of a point atom. The Rayleigh criterion states that for a simple lens with axial illumination a pair of two-dimensional images may be resolved when the maximum of one image is superimposed upon the first minimum of the other, a separation of 0.61 λ (James, 1948). For the case of three-dimensional point atoms, this corresponds to the superposition of the maximum of one image peak on to the first zero of the other, which occurs at an interpeak distance of 0.715 d_{min} (James, 1948). Stenkamp & Jensen (1984) proposed that the distance should be that at which the maximum of the first peak is superimposed with the first minimum of the second, which is equal to 0.917 d_{min}. They added that for atoms with $B > 0$ the image peaks would be broadened, and the local minimum between two peaks at a given separation would therefore become shallower.

Swanson (1988) considered the case for real atoms, with $B > 0$. The effective resolution, D_{eff} is determined, at low resolution, by the high resolution limit and at atomic resolution, 1.2 - 0.8 Å there is a limiting constant value of D_{eff}, reflecting the nature of the atomic peaks. At intermediate resolution, both factors play a part. The intermediate range, 1 - 2 Å, corresponds to that of high resolution protein crystallography. The changing visibility of atoms across this resolution range is a key issue in high resolution protein crystallography. In this range lies the point at which anisotropic refinement of a structure becomes beneficial (Dauter *et al.*, 1995) and the resolution at which structure solution by direct methods may be possible (Sheldrick *et al.*, 1993). At ultra-high resolution, a further level of detail becomes apparent in the density map, corresponding to the deformation density from the spherical atom approximation, but charge density studies on macromolecules will not be feasible in the near future.

Swanson proposed that D_{eff} be defined by a separation distance related to the distance between the maximum of an average image peak and its first inflection point. This was justified because, as B increases, the peak shape changes and the first minimum becomes shallower and moves further from the central maximum at a faster rate than the first inflection point. The average image peak was generated from the Fourier transform of the resolution dependent average structure factor, calculated by averaging data in resolution bins. The first inflection point occurs at the first zero of the second derivative of the density function. For the example structures given, the effective resolution is found to be in the range 1.1 - 1.2 d_{min}.

The Shape of Electron Density

The resolvability of atoms was assessed by direct inspection of density maps. This is justified since it is the resolvability of atoms in the actual density map which affects the course of refinement. While it is true that thermal motion results in the broadening and merging of peaks, this effect can be at least partially compensated for, as described above, by using sharpened maps.

The visibility of atoms in the density for maps at different resolutions and with different degrees of sharpening was investigated by examining the shape of the density between neighbouring atomic centres. All pairs of fully occupied atoms, not including solvent, separated by distances of 1.9 Å or less were selected. The electron density at the two atomic centres, and at nine equally spaced points along the line connecting them, was calculated. The symmetrically placed pairs of points which were equidistant from the midpoint were averaged. The resulting values for the density between each pair of atoms

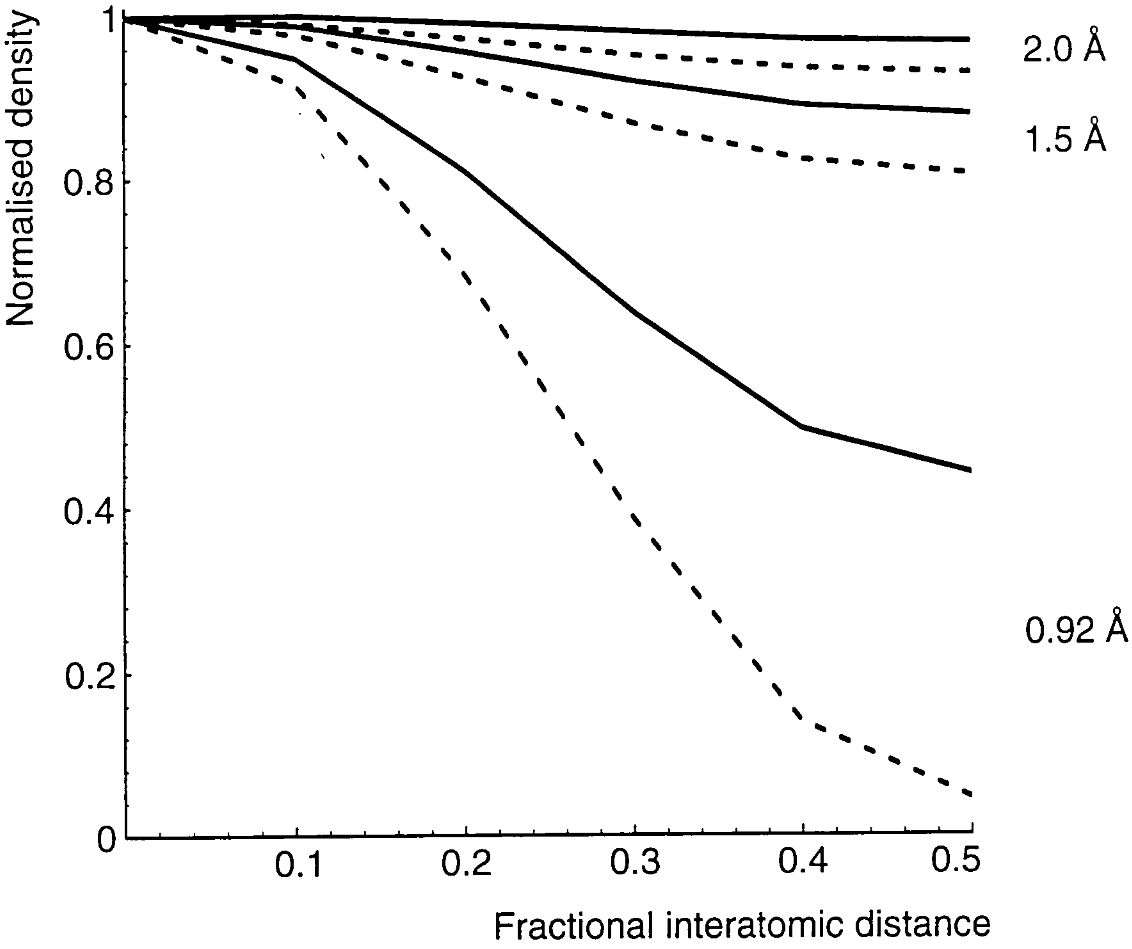


Figure 2. Average shape of the density between neighbouring atoms.
Maps of rubredoxin using data of different sharpness and resolution were investigated. The normalised density values are plotted against fractional distance from the atomic centre, for E-maps (dashed line) and F-maps (continuous line) for maps with 0.92, 1.5 and 2.0 Angstrom high resolution data cutoffs applied.

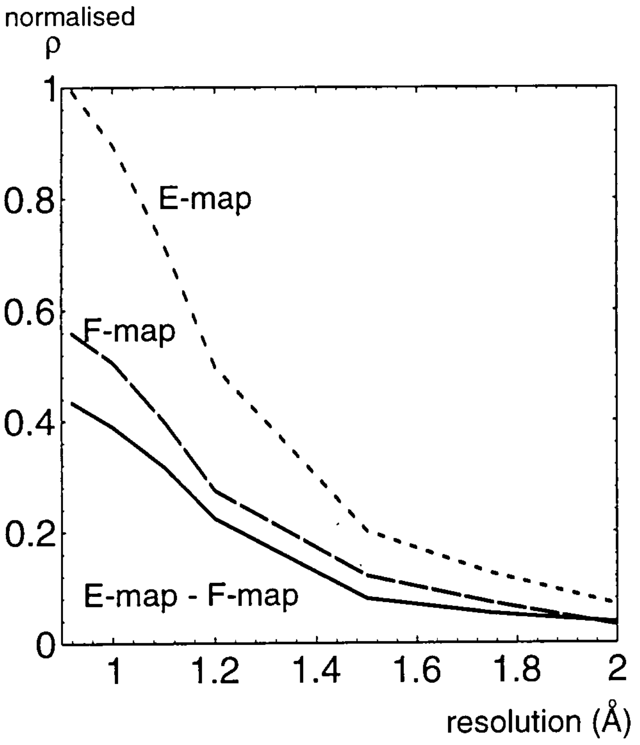


Figure 3.

A comparison of the resolvability of atoms on E maps and F maps at different resolutions for rubredoxin.

If ρ_{atom} = averaged, normalised density at the atom centre and

$\rho_{midpoint}$=density at the midpoint, calculated as explained for Figure 2, then

$\rho_{difference} = \rho_{atom} - \rho_{midpoint}$

$\rho_{difference}$is plotted against resolution cutoff, for E maps(small dashes), F maps(large dashes) and

$\rho_{difference}$(E map) - $\rho_{difference}$(Fmap) (continuous line).

were normalised to give unity at the atomic centre. These values give a representation of the average shape of the density between neighbouring atoms, Figure 2.

If ρ_{atom} is the averaged, normalised density at the atom centre, $\rho_{midpoint}$ is the density at the midpoint between two atoms and $\rho_{difference}$ is the difference between ρ_{atom} and $\rho_{midpoint}$, then the atoms can be said to be 100% resolved if

$\rho_{atom} = 1$ and $\rho_{midpoint} = 0$ therefore $\rho_{difference} = 1$.

The plot of $\rho_{difference}$ against resolution cutoff, Figure 3, shows that the atoms on the E-map are around 40% better defined at atomic resolution, i.e

$\rho_{difference}$ (E-map) = 1 and $\rho_{difference}$ (E-map) - $\rho_{difference}$(F-map) = 0.4

The resolvability of atoms in maps of all degrees of sharpness declines as the high angle data are cut. However, the extra interpretability of E-maps over F-maps increases. At 2.0 Å the resolvability of atoms is 8% in E-maps, 4% in F-maps, which is a 100% advantage for the E-maps. An advantage is retained until around 2.25 Å to 2.5 Å.

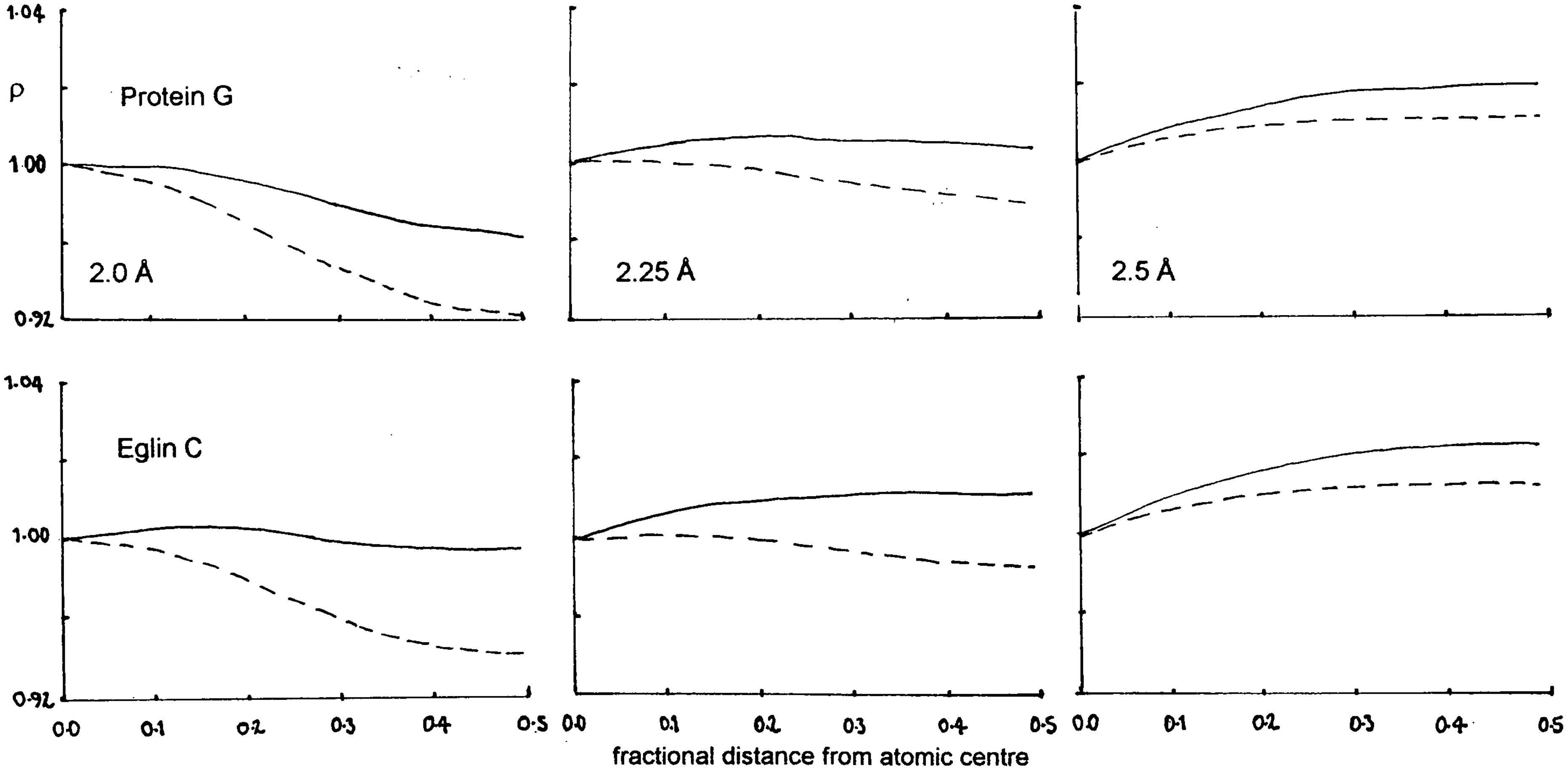
The shape of the electron density between atoms is principally determined by the resolution of the data. At atomic resolution, there is a pronounced minimum at the midpoint, Figure 2. This minimum is lower in the E-map. The depth of the minimum decreases, for all maps, as the resolution is cut. At around 2.5 Å the shape of the density is completely altered and there is a maximum at the midpoint, Figure 4. Between 2 Å and 2.5 Å it may be possible to resolve atoms on the E-map, as the minimum is still present, although it is not in the F-map.

The shape of the density on the F-map is influenced by the thermal parameters of the structure. For a structure with higher B factors, the atoms are less resolved. A map in

Figure 4,

A comparison of the shape of interatomic density for protein G and eglin C.

The normalised average density on maps at six points between the atom centre and midpoint were computed as explained for Figure 2 and plotted against fractional distance from the atom centre. The top row of plots are for protein G, the bottom row, eglin C. From left to right, the highest resolution was 2.0 Å, 2.25 Å and 2.5 Å. Plots are for E-maps (dashes), F-maps (continuous line).



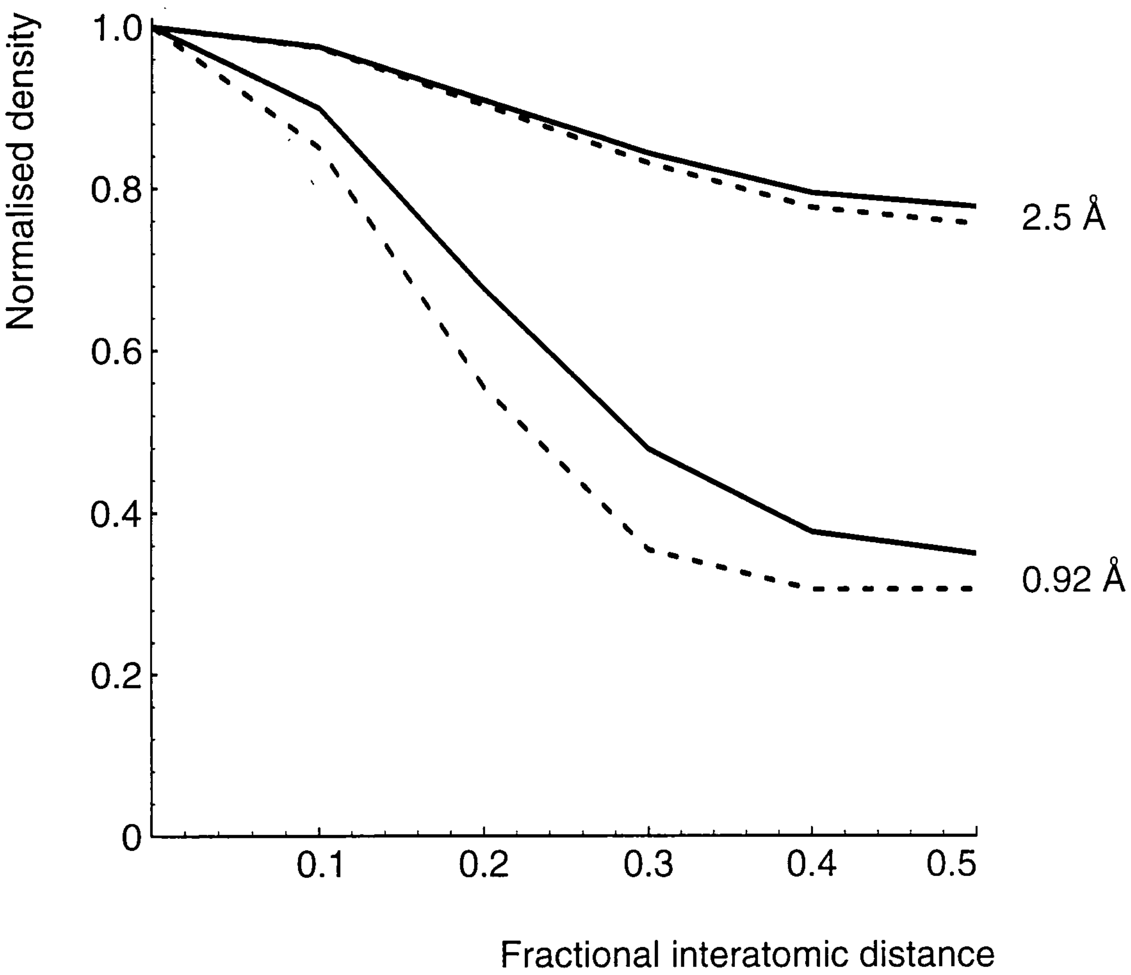


Figure 5a. The resolvability of solvent atoms.

E-map (dashed line) and F-map (continuous line) density between solvent sites is plotted for rubredoxin at 0.92 and with data cut to 2.5 Angstrom.

which resolution had been artificially cut can be distinguished from one for which data are present up to the diffraction limit. Thus there is a substantial difference between the shape of the density in the F-map for protein G at 2.0 Å, and that for eglin at 2.0 Å, Figure 4, due to the difference in average B factors for the structures, Table 1. Since, ideally, thermal effects should be removed during the calculation of E-values, there is a greater similarity between the shape of the density in the E-maps. Errors are not compensated for so the shape of the E-map densities will not be identical.

The density around solvent atoms was also analysed, Figure 5. The separation between solvent sites is greater than that between adjacent atoms in the protein, so solvent molecules are still distinct at a lower resolution. The change in shape of the density with resolution cutoff was much more gradual for solvent. This demonstrates that the high resolution data contain little information about the solvent.

Electron Density Histograms

If a grid is drawn over the map and a histogram of the (F_o, α_c) electron density at the gridpoints is plotted, this distribution will contain information about the nature of the structure and the correctness of the phases (Cochran, 1952; Podjarny & Yonath, 1977; Main, 1990; Lunin, 1993). The map consists of atomic electron density, which at high resolution is seen as a collection of approximately Gaussian peaks, overlying a random noise contribution. Noise arises from errors in the model, the presence of high and low resolution cutoffs in the data as well as errors in the data. The density histogram of the noise is a Gaussian, centred at zero. The atomic electron density histogram has the shape of a Gaussian rotated through 90°. The range of this function is $0 \leq \rho \leq \rho_{max}$. The overall density histogram adopts a characteristic, skewed shape, resulting from the convolution of two component distributions, with a short tail in the negative density region and a long one for positive density (Main, 1990).

It has been postulated that the best set of phases will result in the most skewed density histogram (Cochran, 1952). This relates to the fact that improving the phases causes a reduction in noise, removing small and negative values of ρ and shortening the $\rho < 0$ tail of the histogram. The skewness of the density histogram reflects the level of noise in the map. If maps are generated using the same phases, but from structure factors with different degrees of sharpening, this discriminator should still apply and the most interpretable map should therefore be that for which the density histogram possesses maximum skewness.

The variation of skewness of the density distribution with map sharpness, for maps of type $(F_o^{x}\ E_o^{(1-x)}, \alpha_c)$, was evaluated, Figure 6. The value of x at which skewness is a maximum is defined as x_{max} , thus the $(F_o^{xmax}\ E_o^{(1-xmax)}\ , \alpha_c)$ map possesses the density distribution of maximum skewness. The plot of x_{max} against resolution cutoff is shown in Figure 7, for all four data sets. According to this plot, at 1 Å resolution, the highly sharpened $(F^{0.2}\ E^{0.8})$ synthesis would be expected to contain the most interpretable features. As the resolution is decreased, the optimal degree of sharpening is reduced. This can be explained by the fact that as more of the high resolution data are removed termination errors increase the noise. Thus a greater 'F' contribution to the structure factors is necessary to dampen the noise, and the F/E balance swings towards F.

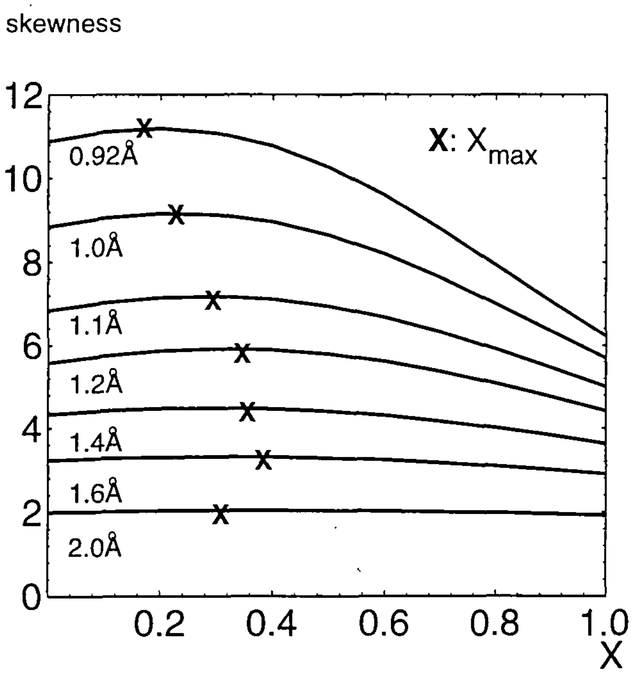


Figure 6.

Skewness of the density distribution was calculated for maps of the form $(Fo^{x} Eo^{(1-x)})$ with x 0->1, for rubredoxin. Plots of skewness against x for maps with a specific high resolution cutoff are shown, for cutoffs at 0.92, 1.0, 1.1, 1.2, 1.6, and 2.0 Angstrom. The value of x at which skewness is maximum, defined as x_{max}, is marked by **X** for each plot.

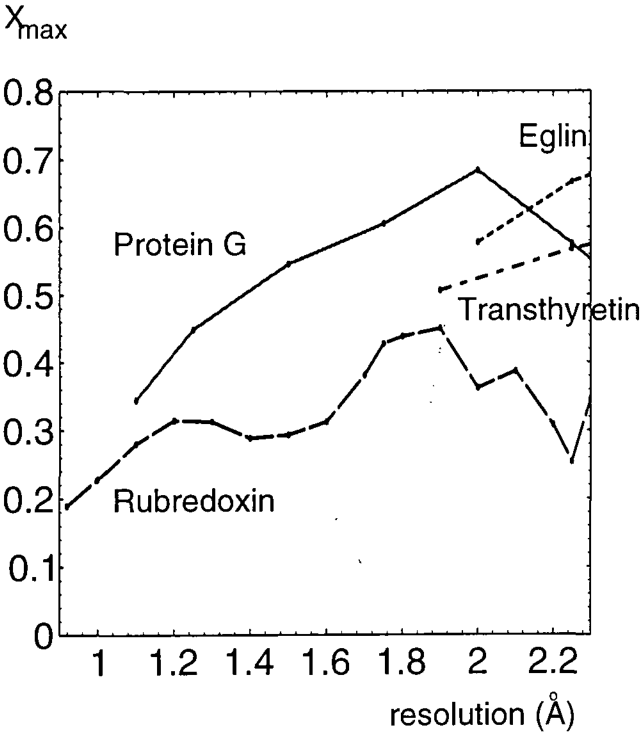


Figure 7.

The plot of x_{max}against resolution cutoff is shown for rubredoxin(large dashes), protein G (continuous line), transthyretin(medium dashes), and eglin C(small dashes).

x_{max} is the value of x at which the skewness of (Fo ^{x}Eo $^{(1-x)}$) maps is maximum.

In other words, the (Fo^{xmax}Eo $^{(1-xmax)}$) map possesses the density distribution of maximum skewness

Refinement using Sharpened Maps

Use of maps of varying degrees of sharpness during real-space improvement steps of refinement has been investigated. A structure obtained from the anisotropic refinement of rubredoxin from *Desulfovibrio vulgaris* at 0.9 Å and five other models derived from this first model were subject to refinement.

Refinement

Restrained least-squares refinement of atomic positions and thermal parameters was performed. The two models referred below to as I and II were refined using SHELXL-93 (Sheldrick, 1993), the remaining examples using PROLSQ. The Automated Refinement Procedure (ARP, Lamzin & Wilson, 1993) was employed for modification of the structures in real space. The refinements were repeated with variation in the sharpness of the $(3F_o\text{-}2F_c,\alpha_c)$ map input to ARP: unsharpened (F), fully sharpened (E) and half-sharpened ($F^{0.5}E^{0.5} \equiv H$) structure factors were used in turn. An account of general refinement protocol is given in the introduction to Chapter 4.

The following models were refined;

I: An anisotropic model, comprising residues 1-52, with 82 water molecules and R and R_{free} values of 8.3% and 11.2%. The SHELXL-93 diffuse solvent correction, based on Babinet's principle (Langridge *et al.*, 1960), had been applied during the previous cycles of refinement. The resolution range of the data was 20 Å to 0.92 Å. 5% of the reflections had been removed from the working dataset for calculation of R_{free}.

II: Solvent atoms with B factor > 30 Å^2 were removed from ***I***, leaving 32 waters. R and R_{free} values were 10.4% and 12.2% respectively.

III: A random positional error with rms 0.3 Å was introduced into model ***I*** and, in addition, the co-ordinates were shifted by 0.5 Å along the *a* axis. This mimics inaccuracies which could be present in a model obtained by molecular replacement. R and R_{free} were 44.8% and 45.9%.

IV: Molecular replacement was carried out using AMORE (Navaza, 1994). The search model used was the 1.4 Å structure of rubredoxin from *Desulfovibrio gigas* (Frey *et al.*, 1987), which has an rms displacement of 0.65 Å from ***I***, for CA atoms. There are 14 sequence differences between the two structures. Side chain shortening mutations only were carried out, where this was appropriate. 7 residues in the molecular replacement model were mutated to Ala and 2 to Ser, leaving 9 sequence differences between the starting model and ***I***. R and R_{free} were 39.2% and 41.0%.

V: A loop region in ***I***, comprising residues Pro 20 to Val 24, was removed. R and R_{free} were 20.9% and 24.3%.

For ***I*** & ***II***, five cycles of SHELXL-93 anisotropic refinement were run, followed by a cycle of ARP. This was iterated ten times. ARP was used for modification of solvent only. The distance limits for addition of new atoms were set to 2.2-3.3 Å and the merging distance to 0.6 Å. Refinement was run with and without application of the SHELXL-93 diffuse solvent correction and with and without real space refinement. The number of atoms to be added and removed in each cycle was set to 0 or 5 for ***I*** and 10 or 15, for ***II***.

For ***III***, ***IV*** & ***V***, PROLSQ refinement was performed until further cycles gave no further drop in R factor. Following each cycle, ARP was run. Real space refinement was carried out on all atoms, but only those designated as solvent were cut and added. Atoms were added at distances of 1.0-3.3 Å from existing ones, and merged if they came within 0.6

Å. These limits were set to allow for the fact that some of the 'solvent' may represent protein atoms. The change in R factor was useful in determining the convergence point of a refinement, while the refinement parameters were being tuned. However, a comparison of R factor does not give a good assessment of how well the refinement process corrects the deliberate errors introduced, since R factors refer to the whole model.

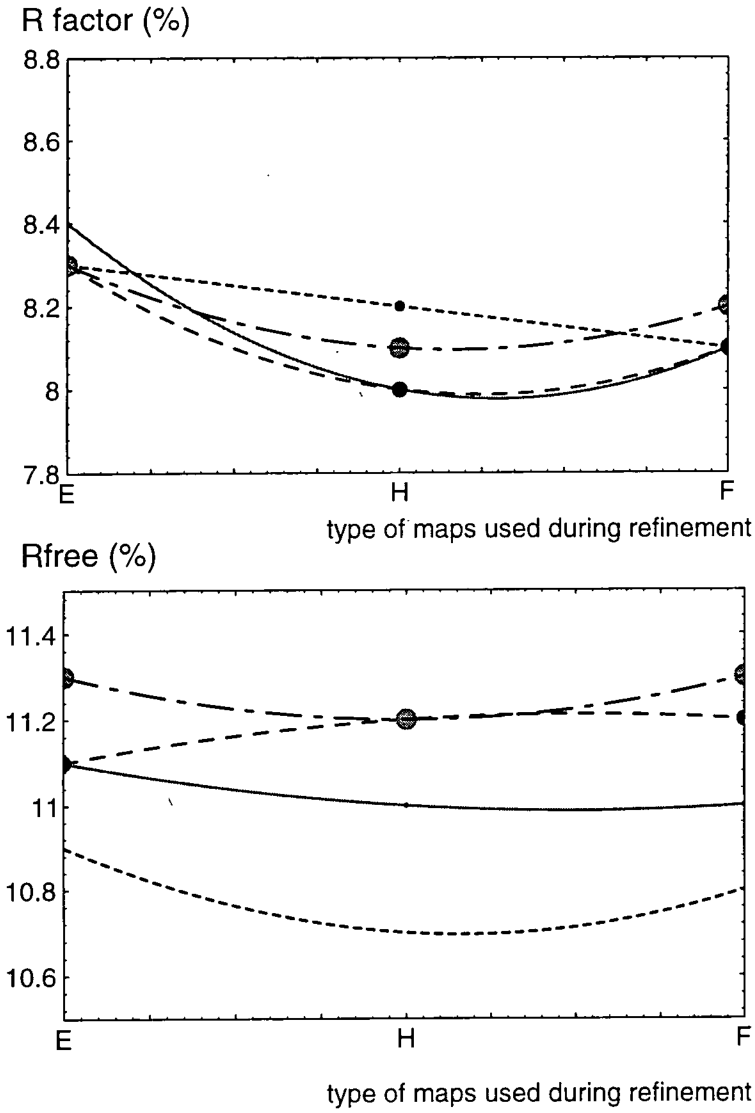
Results of Refinement

In the final stages of refinement, the well defined part of the model remains virtually unchanged, while improvements are made in the fitting in the disordered regions and solvent. Such was the case for the refinements of ***I*** and ***II***. Since scattering from the regions which were modified by these refinements does not contribute greatly to the high resolution data, the effect of sharpening was not dramatic.

Final R and R_{free} values, Figures 8*a* and *b* are almost uniformly lowest for refinements performed using H maps. Use of H maps is advantageous, although the effect is small. Use of fully sharpened maps was ineffective, causing the maximum number of atoms to be removed and added on each cycle, arguing that the noise level in these maps was too high, while semi-sharpened maps were more useful than plain F-maps. Real space refinement assisted in the equilibration of the solvent building during the reconstruction of the solvent network of model ***II***. This can be explained by the observation that diffuse atoms tend to drift towards the edges of the density, a problem corrected for by real space fitting.

Figure 8a. Results of test refinements of rubredoxin.

Final R and Rfree from the refinement of model I is plotted against sharpness of maps used in the refinement; E, H, and F maps.



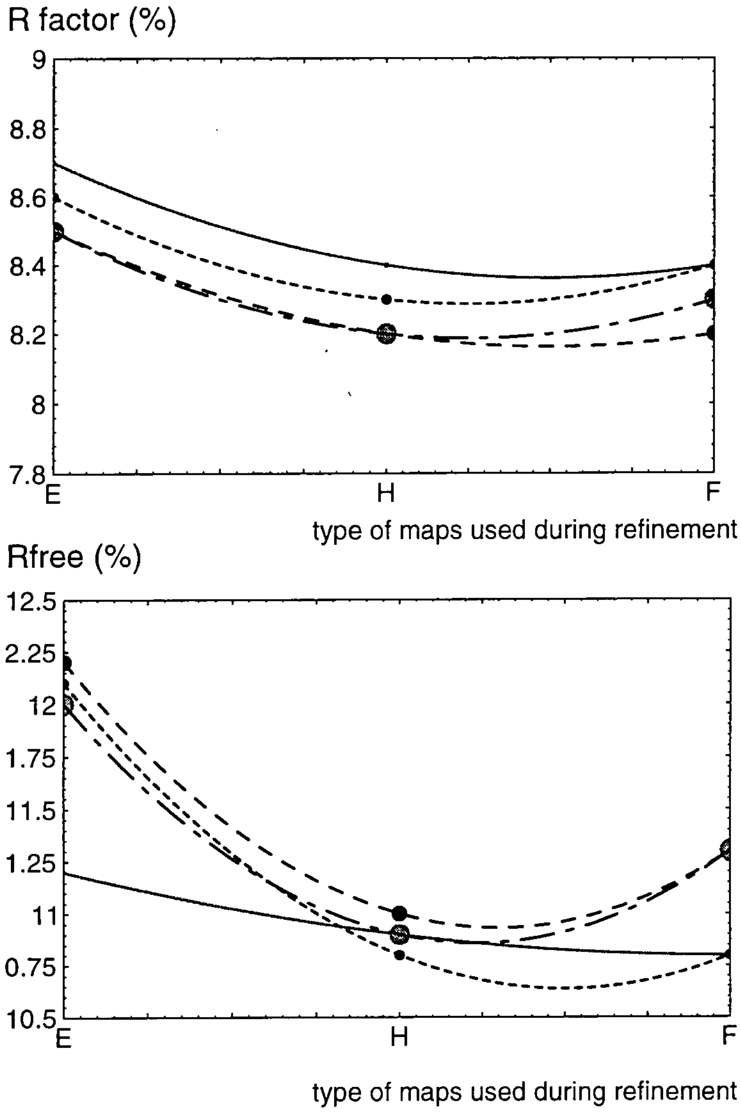
Refinement conditions:

RS(continuous line), R(large dashes), S(small dashes) and neither(dash-dot).

R = real space refinement performed, S = diffuse solvent correction applied.

Figure 8b. Results of test refinements of rubredoxin.

Final R and Rfree from the refinement of model II is plotted against sharpness of maps used in the refinement; E, H, and F maps.

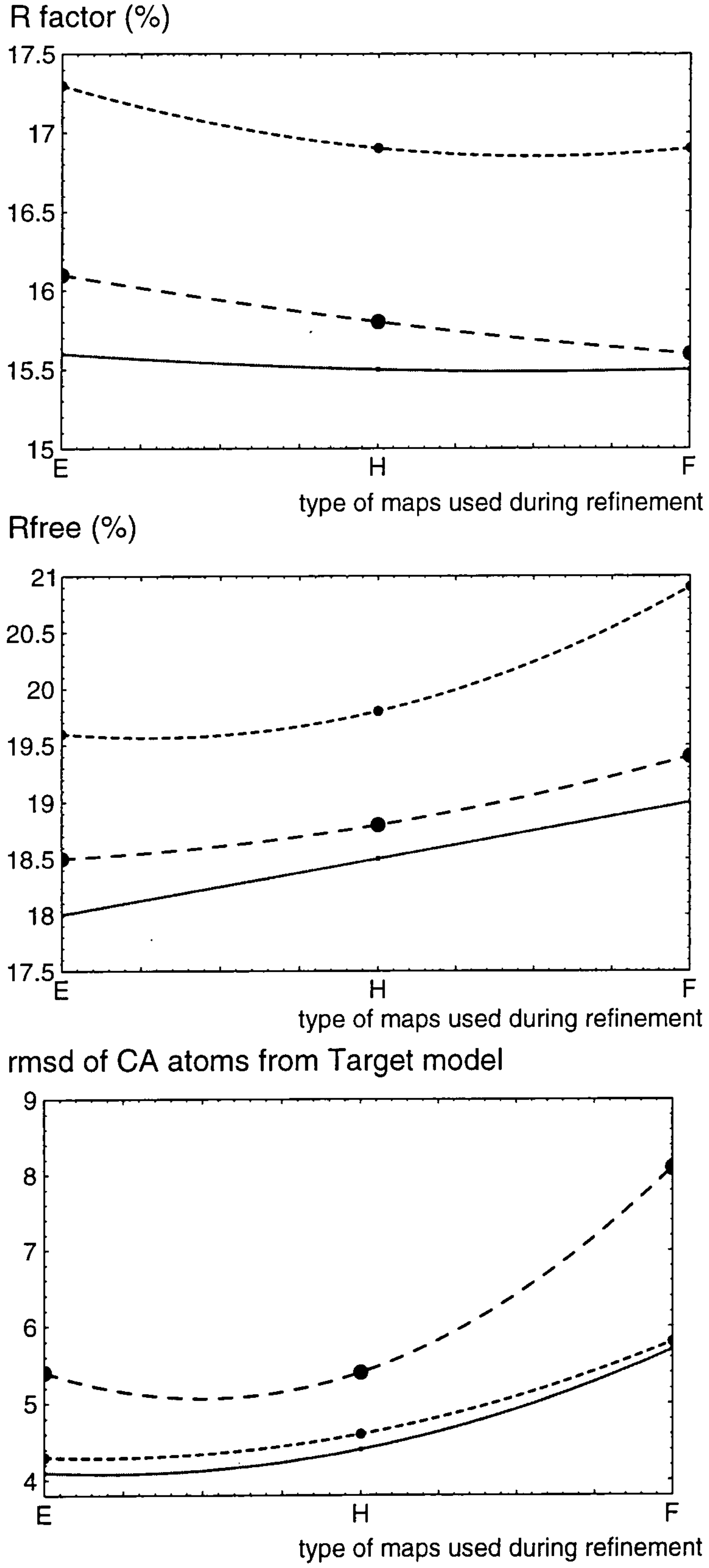


Refinement conditions:

RS(continuous line), R(large dashes), S(small dashes) and neither(dash-dot).

R = real space refinement performed, S = diffuse solvent correction applied.

Figure 8c. Final R, Rfree and rmsd of CA atoms from those in the target model, I are plotted for refinement of models III, IV, and V using E, H, and F maps. model III(continuous line), model IV(large dashes) and model V(small dashes)



The effect of the application of the SHELXL-93 diffuse solvent correction was much more noticeable, since this correction is specific to the low resolution data. When the solvent correction was implemented during the building of a very incomplete solvent network, the addition of solvent was slowed down. When the virtually complete model ***I*** was refined, with the solvent correction operative, the resulting model had fewer solvent molecules and a lower value of R_{free}, Figure 8*b*. This can be ascribed to the removal of peaks which were present due to incorrect scaling of low resolution terms.

Models ***III***, ***IV*** and ***V*** roughly approximate to structures at earlier stages in refinement, with significant errors in the well defined part of the density. The degree to which refinement has corrected the inaccuracies which were introduced can be assessed by observing change in the rms displacement of the main chain atoms from those of model ***I***, Figure 8*c*. ***IV***, the molecular replacement model, possesses 3 regions in which the position of the chain is seriously in error and requires interactive graphical rebuilding. When these regions are not used in the calculation of rms displacement of CA atoms from those ***I***, the values obtained closely mirror those found for model ***III***. R_{free} values reflect the success in correcting the mistake in the main chain, while R values are insensitive. In all three cases, refinement using E-maps produced the best final model. The results from using H-maps were similar to those obtained with E-maps, while the F-map based models were considerably worse.

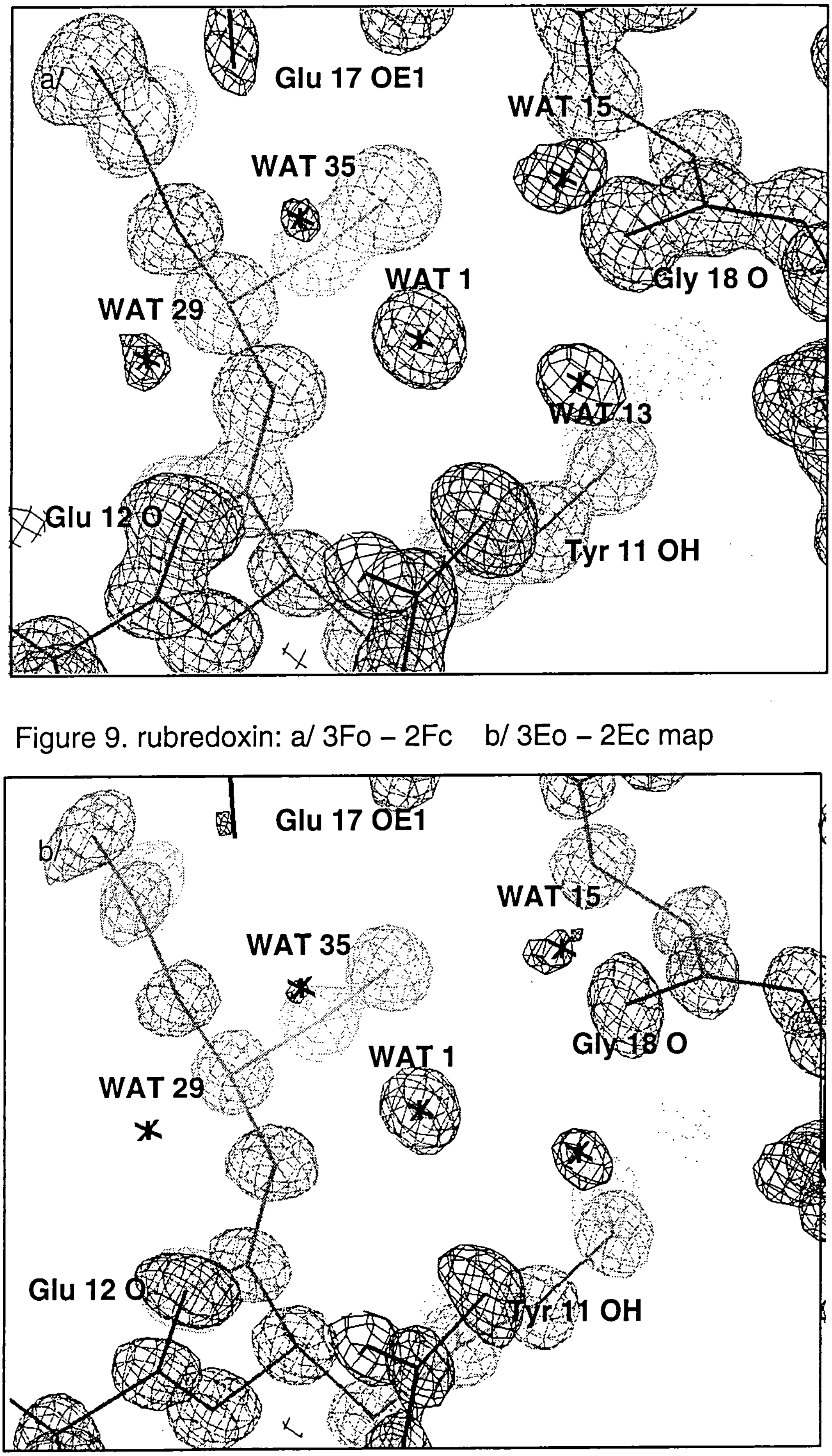


Figure 9. rubredoxin: a/ 3Fo – 2Fc b/ 3Eo – 2Ec map

Conclusions

This investigation was concerned with the resolvability of atoms in maps of different sharpness. Plots of the mean density between adjacent atoms show that fully sharpened maps contain the best resolved peaks in well resolved parts of the structure, due to the removal of peak broadening thermal effects. Concurrent with the increasing sharpness of peaks in the protein region, there is an increase in the background noise level, and a consequent reduction in peak size in less well defined regions, as is illustrated by the density maps of rubredoxin in Figure 9.

The interplay of these factors is resolution dependent. If maximum skewness in the density histogram corresponds to minimum noise in the map, the optimal level of sharpness increases with resolution: around ($E^{0.8}F^{0.2}$) at atomic resolution and between ($E^{0.4}F^{0.6}$) and ($E^{0.6}F^{0.4}$) at 2 Å resolution.

The results of refinements performed on rubredoxin demonstrate the effectiveness of sharpening in the early stages, while showing that this approach is less advantageous towards the end of a refinement. For the atomic resolution rubredoxin structure, the optimal degree of sharpening appears to lie between fully (E) and half-sharpened ($E^{0.5}F^{0.5}$) maps. For an almost fully refined model, the use of half-sharpened maps is most effective, while the worse the model, the sharper the maps should be. These results agree with the trend shown in minimum map noise levels by the density histogram study, Figure 7.

Use of sharpening during automated improvement of the model in real space directly enhances the accuracy with which atoms can be placed within well resolved density.

This can be seen as the primary effect of sharpening. Where significant errors are present in well-defined regions, improvement in the model is accelerated and enhanced. A secondary effect of sharpening stems from the reduction in noise in the density map which results from this improvement in the model. Weaker peaks emerge above the background noise level, allowing effective modelling of more peripheral parts of the model, such as mobile side chains and the solvent network. In the closing stages of refinement, as most of the atoms are already well positioned, potential improvements due to map sharpening are small.

The level of sharpness which produces the most informative map is strongly resolution dependent. While sharpening is most effective for atomic resolution data, it should be advantageous for any refinement at a resolution higher than 2.5 Å, a range which encompasses more than half the structures at present in the Protein Databank (Lamzin *et al.*, 1995).

References, Chapter 5A

Betzel, C., Dauter, Z., Genov, N., Lamzin, V., Navaza, J., Schnebli, H.P., Visanji, M. & Wilson, K.S. (1993) Structure of the proteinase inhibitor eglin C with hydrolysed reactive centre at 2.0 Å resolution. *FEBS Lett.* **317**, 185-188.

Blessing, R.H. & Langs, D.A. (1988) *A priori* estimation of scale and overall anisotropic temperature factors from the Patterson origin peak. *Acta Crystallogr.* **A44**, 729-735.

Blessing, R.H. & Langs, D.A. (1996) Statistical expectation value of the Debye-Waller factor and E(*hkl*) values for macromolecular crystals. *Acta Crystallogr.* **D52**, 257-266.

Cochran, W. (1952) A relation between the signs of structure factors. *Acta Crystallogr.* **5**, 65-67.

Damas, A.M., Ribeiro, S., Lamzin, V.L., Porto, J.A. & Saraiva, M.J. (1996) The crystal structure of Val-122-Ile variant transthyretin - a cardiomyopathic mutant. *Acta Crystallogr. D.* in press.

Dauter, Z., Sieker, L.C. & Wilson, K.S. (1992) Refinement of Rubredoxin from *Desulfovibrio vulgaris* at 1 Å with and without restraints. *Acta Crystallogr.* **B48**, 42.

Dauter, Z., Lamzin, V.L. & Wilson, K.S. (1995) Proteins at atomic resolution. *Curr. Opin. Struct. Biol.* **5**, 784-790.

Derrick, J.P. & Wigley, D.B. (1994) The third IgG-binding domain from streptococcal protein G. An analysis by X-ray crystallography of the structure alone and in a complex with Fab. *J. Mol. Biol.* **243**, 906-918.

Frey, M. Sieker, L.C., Payan, F., Haser, R., Bruschi, M., Pepe, G. & Le Gall, J. (1987) Rubredoxin from *Desulfovibrio gigas.* A molecular model of the oxidised form at 1.4 Å. *J. Mol. Biol.* **197**, 525.

James, R.W., (1948) False detail in three-dimensional Fourier representations of crystal structures. *Acta Crystallogr.* **1**, 132-134.

Karle, J. and Hauptmann, H. (1953) Application of statistical methods to the naphtalene structure. *Acta Crystallogr.* **6** 473-476.

Karle, J. and Hauptmann, H. (1956) A theory of phase determination for the four types of non-centrosymmetric space grous *1P222, 2P22,* $3P2_12$, $3P_22$. *Acta Crystallogr.* **9**, 635-651.

Lamzin, V.S., Sevcik, J., Dauter, Z. & Wilson, K.S. (1995) Implications of atomic resolution. *Making the most of your model. Proceedings of the CCP4 Study Weekend, 6-7 January, SERC Daresbury Laboratory, Daresbury, Warrington, England,* 33-40.

Langridge, R., Marvin, D.A., Seeds, W.E., Wilson, H.R., Hooper, C.W., Wilkins, M.H.F. & Hamilton, L.D. (1960) The molecular configuration of Deoxyribonucleic Acid. II. Molecular models and their Fourier transforms. *J. Mol. Biol.* **2**, 38-64.

Lunin, V.Y. (1993) Electron-density histograms and the phase problem. *Acta Crystallogr.* **D49**, 90-99.

Main, P. (1990) A formula for electron density histograms for equal-atom structures. *Acta Crystallogr.* **A46**, 507-509.

Podjarny, A.D. & Yonath, A. (1977) Use of matrix direct methods for low-resolution phase extension for tRNA. *Acta Crystallogr.* **A33**, 655-661.

Sheldrick, G.M., Dauter, Z., Wilson, K.S., Hope, H. & Sieker, L.C. (1993) The application of direct methods and Patterson interpretation to high-resolution protein data. *Acta Crystallogr.* **D49**, 18-23.

Stenkamp, R.E. & Jensen, L.H. (1984) Resolution revisited: limit of detail in electron density maps. *Acta Crystallogr.* **A40**, 251-254.

Swanson, S.M., (1988) Effective resolution of macromolecular X-ray diffraction data. *Acta Crystallogr.* **A44**, 437-442.

Tronrud, D.E. (1996) The TNT Refinement Package. In *Methods in Enzymology* (Carter, C. & Sweet, B. eds.) in press.

Wilson, A.J.C. (1942) Determination of absolute from relative X-ray data intensities. *Nature* **150**, 151-152

see also general crystallographic references, end of Chapter 5

Chapter 5B

Treatment of solvent in protein crystal structures

List of abbreviations used in the description of solvent structure, Chapter 5B

V_m	crystal volume per unit molecular weight (Matthews, 1968) ($Å^3$ D^{-1})
V_{ASSY}	volume of asymmetric unit ($Å^3$)
ρ_{ASSY}	average density of asymmetric unit (g cm^{-3})
V_{prot}, $V_{f,prot}$	volume of protein in asymmetric unit ($Å^3$), fractional volume of protein
M_{prot}	mass of protein in asymmetric unit
ρ_{prot}	average density of protein (g cm^{-3})
V_{sc}	% volume of solvent by relative cell and protein volumes, from equation (3).
$\rho_{solvent}$	average density of solvent (g cm^{-3})
$N_{sol,modelled}$	number of solvent sites modelled in asymmetric unit
$N_{sol,total}$	total number of solvent sites, calculated from V_{sc}, with the approximation $\rho_{solvent} = 1$
$V_{sol,ordered}$	% volume of ordered solvent, from proportion total:modelled solvent, equation (9)
$V_{sol,diffuse}$	% volume of diffuse solvent, from proportion total:modelled solvent, equation (9)
V_{ss}	% volume of ordered solvent, from analysis of gridpoint-atom separation, Figure 5
V_{ra}	% volume remote from protein, from analysis of gridpoint-atom separation, Figure 5
V_{ds}	% volume of diffuse solvent, from analysis of gridpoint-atom separation, Figure 5
k_{sol}	contrast parameter for the modelling of diffuse solvent, defined as $\Sigma\rho_{sol}(r)$ / $\Sigma\rho_{protein}(r)$.
B_{sol}	average B factor for diffuse solvent ($Å^3$)
d	resolution (Å)
s	resolution expressed as $\sin\theta / \lambda$ ($Å^{-1}$)

Introduction

What is solvent structure?

The essential difference between small molecule crystals and those of macromolecules (although a sharp boundary does not exist, a continuum of examples exists between large and small) is that a large part of the volume of the latter is taken up by solvent. Realistic modelling of this solvent is a matter of concern because of the fundamental nature of a crystallographic model. Model phases are generated by combining the scattering contributions from all the constituents of the unit cell. Consequently, the quality of the whole model affects the detail which can be seen in each part and careful modelling of the solvent is necessary to achieve the most informative model of, for example, the active site.

A measure of packing density is given by V_m, the crystal volume per unit molecular weight (1) (Matthews,1968).

$$V_m = V_{ASSY} / M_{prot} \qquad (1)$$

For protein crystals V_m values lie in the range 1.65 to 3.35 $Å^3Da^{-1}$, most frequently between 2.1 and 2.4 $Å^3Da^{-1}$. The distribution has a sharp cutoff at the lower end, reflecting the fact that the limits of close packing have been reached, while at the upper end there is a long tail, since there is no definite limit to how loosely packed a structure could feasibly be. V_m is related to the fractional solvent content of the crystal (2).

$$V_{f,prot} = V_{prot} / V_{ASSY} = (M_{prot} / \rho_{prot}) (\rho_{ASSY} / M_{ASSY}) = (V_{ASSY}\ \rho_{ASSY}) / (V_m\ \rho_{prot}\ M_{ASSY})$$

$$V_{f,prot} = 1 / V_m\ \rho_{prot} \qquad (2)$$

If V_m is expressed in $Å^3Da^{-1}$ and ρ_{prot} approximated to 1.35 g cm^{-3}, then the solvent content of the crystal can be expressed by (3) (Matthews, 1968).

$$V_{sc} = 100(1-(1.23 / V_m))\ \ \% \qquad (3)$$

The fraction of volume occupied by solvent for a typical protein crystal is between 25% and 65%, with an average of 43%. Although the volume occupied by solvent in a protein crystal varies greatly, the structure adopted by the solvent does not. This justifies the development of a general model for the solvent network and a protocol for solvent treatment which should be valid in most cases.

Ideas about the solvent structure around a protein have been developed from complementary viewpoints; X-ray diffraction crystallography, NMR spectroscopy and theoretical modelling (Karplus & Faerman, 1994; Levitt & Park, 1993). In crystal structures discrete solvent sites are visible, mainly in the boundary region, between protein and diffuse solvent. The number of such sites which are conserved between different crystallographic forms, with the exception of waters buried within the protein, is a subject of contention between NMR spectrocopists and crystallographers.

NMR studies show that, in solution, all except the buried waters are in rapid motion, with exchange rates > $1\text{x}10^{10}$ s^{-1}. Theoretical calculations support this view and suggest that the same is true within the crystal. It may be constructive to consider the solvent sites observed in a crystal structure as the positions of minima in the free energy potential for solvent molecules, influenced as much by crystal field forces as chemical interactions with neighbouring side chains.

The H-bonding network around the surface of the protein incorporates a shell of solvent, which can be modelled as a set of discrete atoms on fully occupied sites. If the resolution of the data permits, partially occupied sites, linked to disordered systems of side chains, may also be included in the model. The remaining solvent is not attached to specific minima. This diffuse solvent contributes to the scattering at low resolution, due to the contrast in electron density between diffuse and ordered regions of the unit cell. There is no sharp boundary between these regions, the density varying smoothly between the centre of the protein molecule and the diffuse solvent zone. Thus, the occupancies of some of molecules in the first solvent shell appear to be less than unity and some discrete solvent peaks may be located further from the molecule, in the essentially diffuse solvent region.

Treatment of diffuse solvent

If the region of a model crystal structure actually filled with diffuse solvent is considered to be a vacuum, there will be systematic errors in the calculated intensities at low resolution, because the contrast in density between protein and diffuse solvent is much smaller than that between protein and an empty void. The scattering of water exhibits periodicity at around 3 Å, but little above 4 Å (Blessing & Langs, 1988). Therefore, at low resolution, its scattering can be considered to be that of a gas of uniform electron density. The volume occupied by the water is equal to the unit cell volume minus the sum of partial atomic volumes of all atoms in the protein model. The partial atomic volume of an atom is similar whether it is surrounded by water or buried in the protein. Babinet's principle states that the scattering, due to water in the space around the molecule, has the same amplitude but opposite phase to the scattering from the molecular volume if it were filled with water (4). If neither scattering pattern has much detail, which is the case at low resolution, and the contrast between the density of water and protein is small, the assumption (5) can be made.

$$\boldsymbol{F}_{\mathrm{sol}}(\mathrm{s}) = -\boldsymbol{F}_{\mathrm{sol\ in\ protein\ volume}}(\mathrm{s}) \qquad (4)$$

$$\boldsymbol{F}_{\mathrm{sol\ in\ protein\ volume}}(\mathrm{s}) \cong \boldsymbol{F}_{\mathrm{protein}}(\mathrm{s}),\ \mathrm{so}\ \boldsymbol{F}_{\mathrm{protein}}(\mathrm{s}) \cong -\boldsymbol{F}_{\mathrm{sol}}(\mathrm{s}) \qquad (5)$$

The contrast between protein and water requires the introduction of a scaling factor, k_{sol}, defined as $<\rho_{sol}(r)> / <\rho_{protein}(r)>$. A resolution dependent factor, $exp(-B_{sol}\, s^2)$, is required to restrict the relation to low angle data. The total scattering from the crystal may be expressed as (6), which leads to the approximation (7). Thus, the introduction of a scaling factor $(1-k_{sol}\, \exp(-B_{sol}\, s^2))$ for F_c's corrects for the diffuse solvent scattering. B_{sol} and k_{sol} may be determined during the scaling of F_c's to F_o's by the minimisation of function (8) (Langridge *et al.*, 1960; Driessen *et al.*, 1989; Tronrud, 1996).

$$\boldsymbol{F}_c(s) = \boldsymbol{F}_{protein}(s) + \boldsymbol{F}_{sol}(s) \qquad (6)$$

$$\boldsymbol{F}_c(s) \cong (1 - k_{sol}\, exp(-B_{sol}\, s^2))\, \boldsymbol{F}_{protein}(s) \qquad (7)$$

$$f = \Sigma\, (\boldsymbol{F}_o(s) - k\, exp(-Bs^2)(1 - k_{sol}\, exp(-B_{sol}\, s^2))\, \boldsymbol{F}_{protein}(s))^2 \qquad (8)$$

Aims

The nature of a diffuse solvent correction of the type described above and the effects of its application during the refinement of a small protein structure have been investigated. For a good assessment to be made of whether the solvent treatment is effective and justified, it is first necessary to have an understanding about the nature of the solvent structure in the crystal.

Experimental

Application of a diffuse solvent correction

Plotting resolution dependence of the diffuse solvent scaling factor, $(1 - k_{sol} exp(B_{sol}\, s^2))$, Figure 1, shows that the presence of diffuse solvent should be accounted for during the refinement of any structure for which there are data at a resolution lower than 5 Å. An alternative is the complete removal of these low resolution data, but this sacrifices structural information and results in extremely noisy density maps, due to truncation errors.

Figure 1.

The diffuse solvent scaling factor (1 - $k_{sol}exp(B_{sol}$ s^2)) is plotted against resolution, 1/d ($Å^{-1}$) This scaling does not effect data of resolution higher than 2.5 Å (1/d > 0.4 Å), while strongly influencing low resolution terms, below 5 Å (1/d < 0.2 Å),

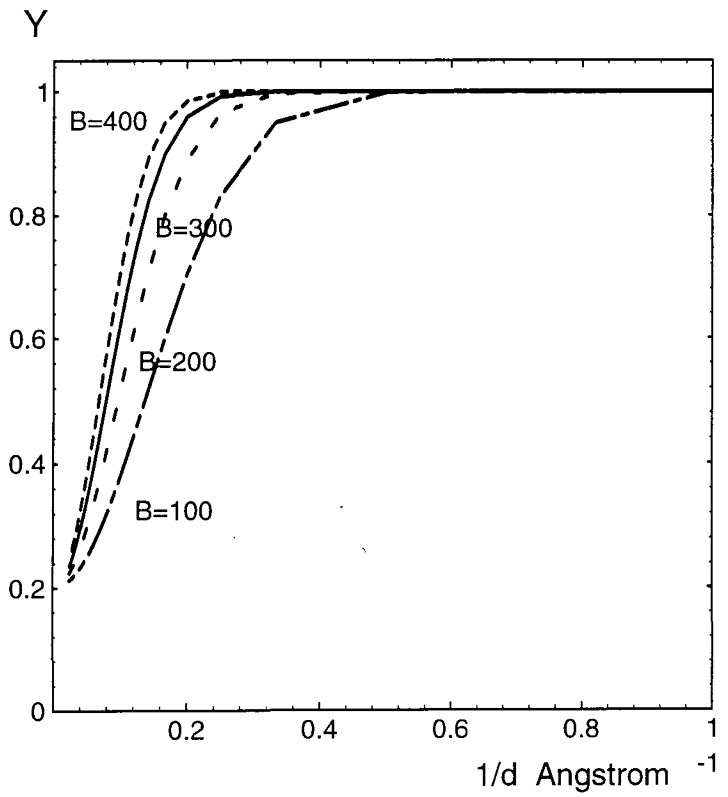


Figure 1.

The diffuse solvent scaling factor $Y=(1-k_{sol}exp(B_{sol}s^2))$ is plotted against resolution, 1/d. This scaling does not effect data of resolution > 2.5 Angstrom (1/d > 0.4), while strongly influencing low resolution terms, below 5 Angstrom (1/d < 0.2).

Values for k_{sol} and B_{sol} must be estimated during refinement, as described above. k_{sol} is the contrast parameter; using an incorrect value is equivalent to setting the wrong average density of diffuse solvent, which results in under or over-scaling of Fc's. B_{sol} determines the manner in which solvent is modelled at the solvent protein interface.

Figure 2*a* illustrates the protein/solvent interface schematically. The effects of setting B_{sol} to reasonable, small and large values are shown. The correct value gives a smoothly changing density across the boundary. If B_{sol} is too small, there will be a 'hole' in the density, if it is too large, there will be a high density 'ridge' at the protein/solvent interface. The resulting ripples in the density may make it impossible to distinguish real features of the model from artificial peaks. In reciprocal space, a badly estimated value of B_{sol} gives rise to an inadequate fit between calculated and observed structure factors at low resolution.

If discrete solvent sites are not modelled, the estimated value of B_{sol} will be small. As solvent is added, B_{sol} increases. When no discrete solvent has been modelled, there is a sharp boundary between protein and diffuse solvent regions. Solvent sites are added at the interface, increasing the width of the threshold region. Thus, a more gradually varying function is necessary to model smooth change in density and the value of B_{sol} increases, as illustrated schematically in Figure 2*b*.

Figure 2*a*. Schematic representation of the protein/solvent interface, illustrating both well and poorly modelled electron density

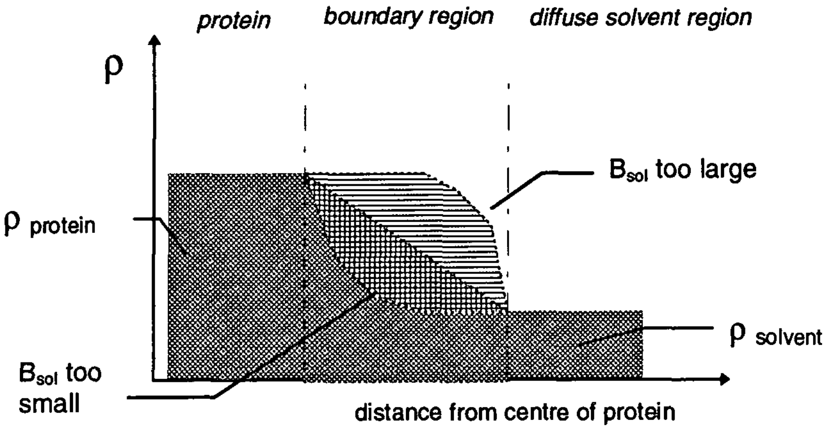
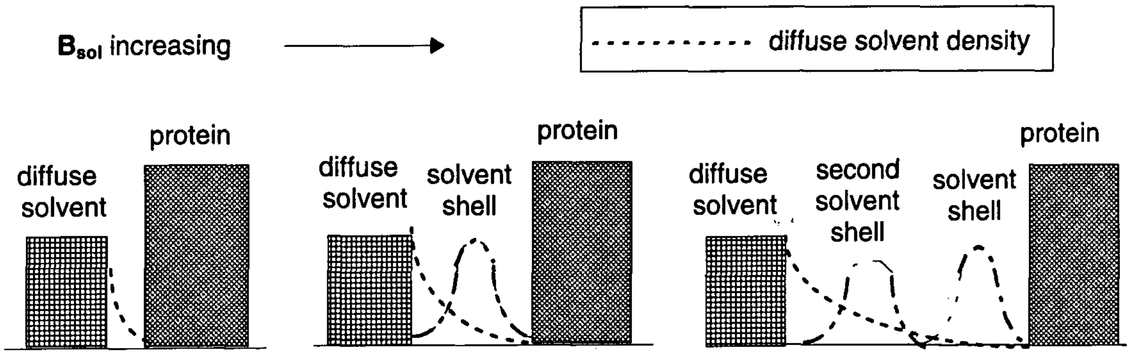


Figure 2*b*.
Schematic Representation of changes in boundary region during construction of a solvent network



(1)
No solvent modelled

(2)
Single shell of solvent modelled around protein

(3)
Second solvent shell modelled

Analysis of the nature of solvent in the crystal structures of two small proteins

The data-to-parameter ratio for a protein crystal structure necessitates the application of numerous restraints on main chain and side chain geometry. This is not required for solvent atoms. Distance constraints may be applied, restricting them to a range of distances from the next atom, so they neither become isolated, nor approach closer than a feasible H-bonding distance, this latter condition being unjustified for overlapping solvent networks. During anisotropic refinement with SHELXL-93 (Sheldrick, 1993) solvent may be restrained to be approximately isotropic, to prevent the volume of the thermal ellipsoid refining to a negative value, which can easily occur, since an ellipsoid is an inadequate model for the density of a typical solvent site. Solvent is otherwise refined without restraints and, as a consequence, the solvent co-ordinate error distribution is an unbiased Gaussian. Thus, analysis of the solvent can give an independent assessment of model quality.

Model protein structures

Two structures were used as examples in this investigation of solvent character; protein G from *Streptococcus* (Derrick & Wigley, 1994, Chapter 4A) and rubredoxin from *Desulfovibrio vulgaris* (Dauter *et al.*, 1992, Chapter 4B), Table 1. Both these models have been subject to anisotropic refinement using SHELXL-93. The models used in this analysis were subsequently refined using all of the data, but at this stage, 5% of the reflections had been kept separate, for the calculation of R_{free}. The occupancy of all solvent sites was 100%.

Table 1. Structural details

	rubredoxin	*protein G*
Space group	$P2_1$	$P2_12_12_1$
Resolution (Å)	20 - 0.92	10.0 - 1.1
B factor from Wilson plot (Å^2)	15	20
mean B factor, protein (Å^2)	9.1	12.5
mean B factor, solvent (Å^2)	32.1	35.3
V_m (Å^{3}Da^{-1})	1.74	2.23
V_{sc} (%)	29	45
R (%)	8.3	9.8
R_{free} (%)	11.2	12.7
$N_{sol,modelled}$	82	119
$N_{sol,total}$	94	222

Application of a diffuse solvent correction during refinement

A suitable approach to the treatment of diffuse solvent during the refinement of protein G was required. Several schemes were tested and their results compared. The refinement process is summarised in Table 2. Poor scaling caused by data collection problems (Chapter 4A) resulted in large errors in the low resolution data. Consequently, a low resolution cutoff of 10 Å was applied. The lack of the lowest angle reflections is not ideal for the modelling of solvent and may result in inaccuracies in calculation and refinement of the diffuse solvent parameters. Large errors and absences in low resolution data are also a source of noise, making the modelling of solvent and disordered side chains in diffuse density more difficult.

Table 2. Refinement of protein G

Stage	
1	isotropic refinement , PROLSQ(Konnert & Hendrickson, 1980)
	+ real space refinement, construction of solvent network, ARP
2	anisotropic refinement , SHELXL-93, ARP
3	introduction of 7 double conformations, occupancies of 8 sidechains refined
4	sharpened ($3F_o$-$2F_c$) maps input to ARP; $F^{0.2}E^{0.8}$

	diffuse solvent scaling				no diffuse solvent scaling			
after Stage	model	R.(%)	Rfree.(%)	total water	model	R.(%)	Rfree.(%)	total water
2	*2A*	10.1	12.9	119	*2R*	10.1	12.9	120
3	*3A*	9.6	12.5	125	*3R*	9.8	12.6	120
4	*4A*	9.5	12.5	132	*4R*	9.7	12.8	120

Anisotropic refinement was performed, using SHELXL-93,[7] with the application of the diffuse solvent correction. The diffuse solvent correction in SHELXL-93 refines the value of k_{sol}, but B_{sol} must be assigned a value which is not refined. In SHELXL-96 (Sheldrick & Schneider, 1996) both k_{sol}, and B_{sol} are refined, but this was a development subsequent to this study. k_{sol} refined to a physically meaningless value, possibly due to truncation errors. The solvent correction was retained, since it appeared to help compensate for these errors, giving the refinement more stability.

During the refinement, the F_o's and F_c's output from SHELXL-93 were scaled, before being used for density map synthesis, providing the input for ARP (Lamzin & Wilson, 1993), which was employed for modification of the structures in real space. Two parallel refinements were performed, the first using the CCP4 program, RSTATS. The second refinement utilised an alternative scaling program, implemented in ARP. This calculates a diffuse solvent correction, as described above, refining values of k_{sol} and B_{sol} (Tronrud, 1996). The refinements were, in all other ways, identical.

The solvent structures of four models, listed in Table 3, were compared. The separation of solvent sites was calculated between pairs of models: *3A* & *3R*, *4R* & *3R*, *4A* & *3A*, Figures 3*a* & *b*. Sites more than 0.5 Å distant from the nearest site in the other model were considered to be different. The positions of these sites in the density were inspected, using FRODO (Jones,1978), with the results shown in Table 4.

Table 3. Models used for comparison of solvent structure treatment

model	result of refinement with sharpened maps input to ARP	result of refinement with more careful diffuse solvent correction applied
3A	no	no
3R	no	yes
4A	yes	no
4R	yes	yes

Table 4. An assessment of the new solvent sites

comparison of / to	new sites	average B (Å^2)	well positioned	badly positioned
3A / 3R	20	65.5	7	10
4A / 3A	14	79.3	5	6
4R / 3R	6	63.5	0	4

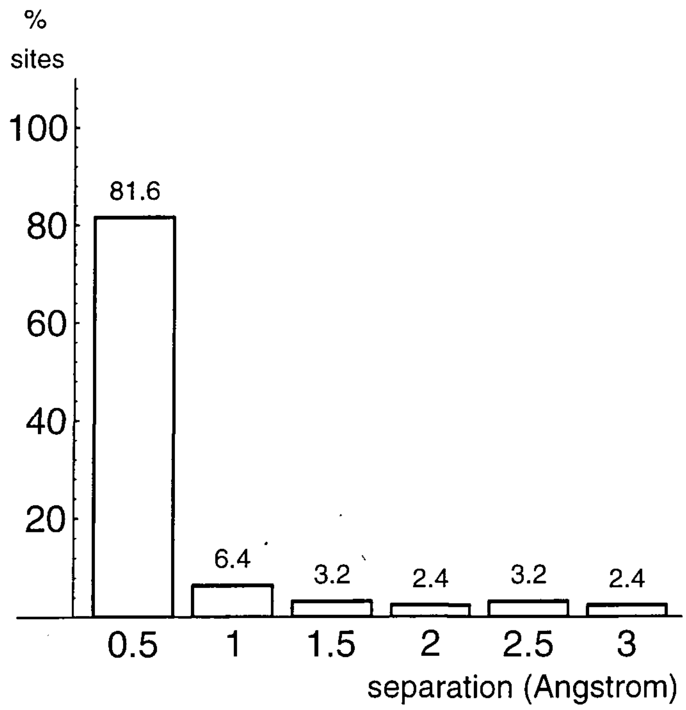


Figure 3.

Histograms of the separation of solvent sites for pairs of protein G models, with nomenclature as defined in Table 3.

(a) Separation of solvent in 3A from 3R, the effect of refinement with diffuse solvent correction applied.

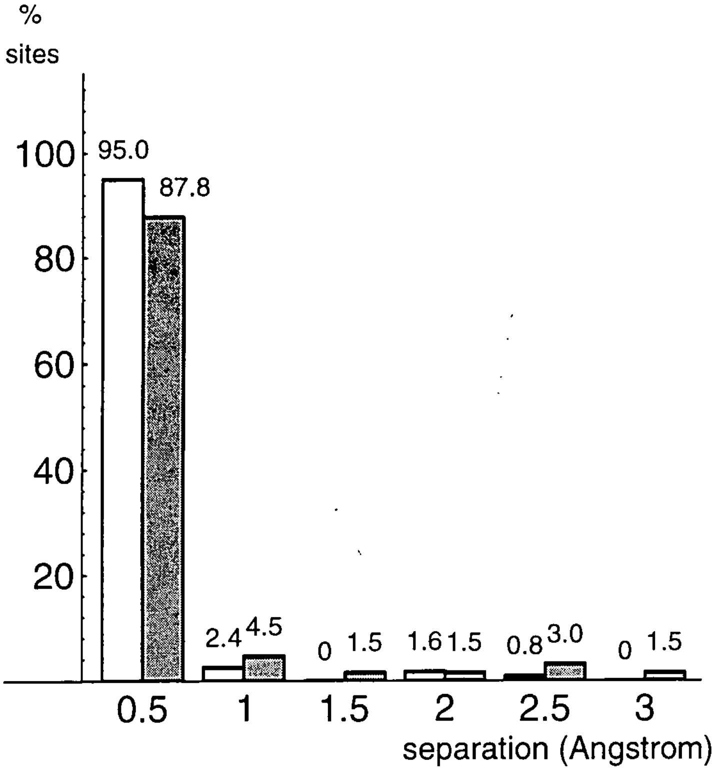


Figure 3.

Histograms of the separation of solvent sites for pairs of protein G models, with nomenclature as defined in Table 3.

(b) Separation of solvent in 4R from 3R(grey) and 4A from 3A(white); the effect of refinement with and without the application of a diffuse solvent correction.

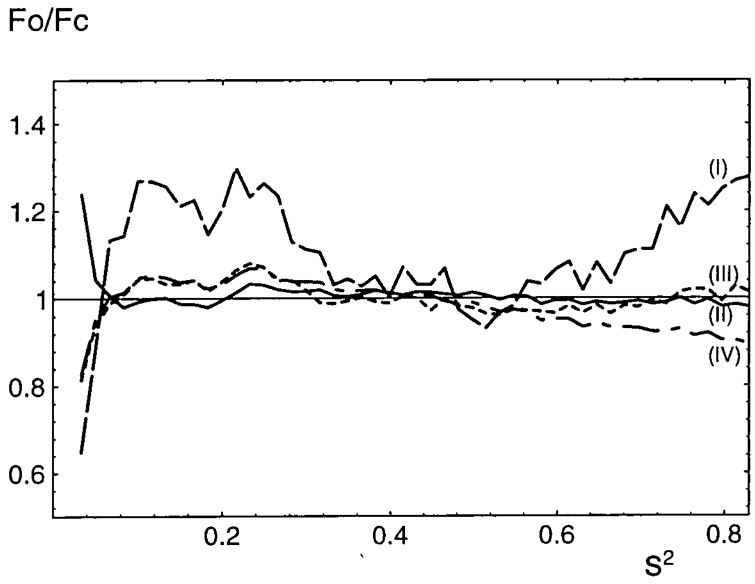


Figure 4.

An assessment of the match of calculated and observed structure factors. Fo and Fc for protein G models at sequential stages in refinement, with nomenclature as defined in Table 5, were sorted into resolution bins. The fit of Fo and Fc at different stages in refinement was compared.

(a) Fo/Fc against $\sin^2 \theta / \lambda^2$

model I(large dashes), model II(small dashes), model III(dash-dot), model IV(continuous line).

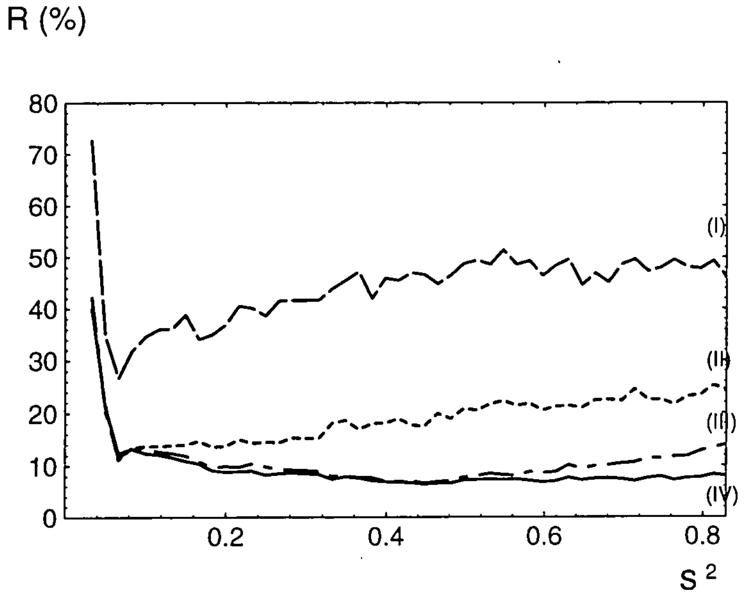


Figure 4.

An assessment of the match of calculated and observed structure factors. Fo and Fc for protein G models at sequential stages in refinement, with nomenclature as defined in Table 5, were sorted into resolution bins. The fit of Fo and Fc at different stages in refinement was compared.

(b) R against $\sin^2 \theta / \lambda^2$

model I(large dashes), model II(small dashes), model III(dash-dot), model IV(continuous line).

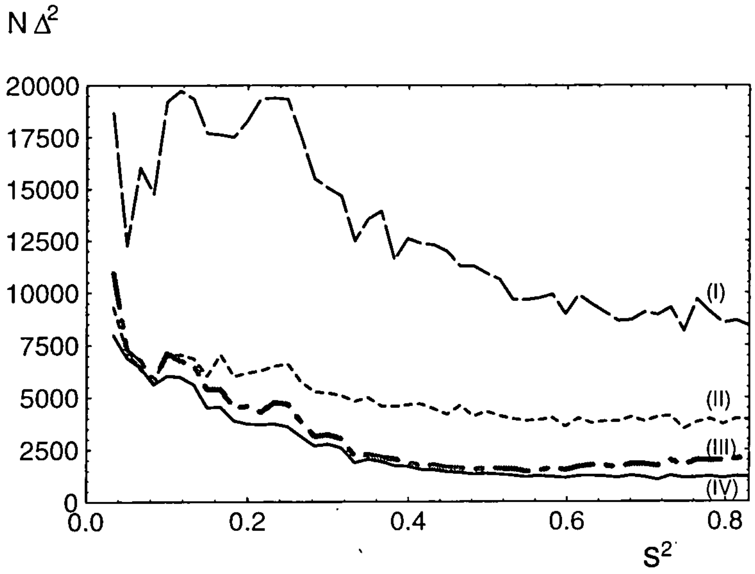


Figure 4.

An assessment of the match of calculated and observed structure factors. Fo and Fc for protein G models at sequential stages in refinement, with nomenclature as defined in Table 5, were sorted into resolution bins. The fit of Fo and Fc at different stages in refinement was compared.

(c) N Δ^2 against sin $^2\theta / \lambda^{\ 2}$

model I(large dashes), model II(small dashes), model III(dash-dot), model IV(continuous line).

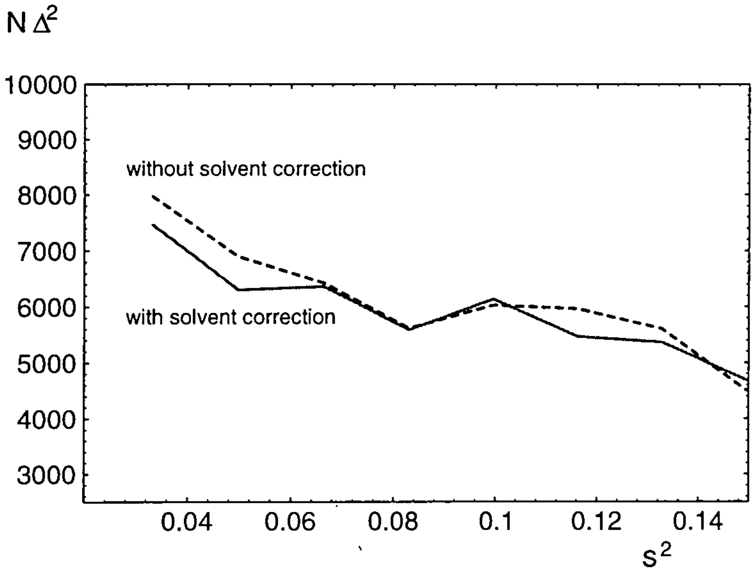


Figure 4.

The difference in fitting at low resolution resulting from the application of the diffuse solvent correction is too small to show up on the scale of (a), (b), and (c)

(d) N Δ^2 against sin $^2\theta / \lambda$ ^{2}

model IV(continuous line) and model V(dashes).

Application of the diffuse solvent correction lead to a decrease in the R factor of 0.19% and in R_{free}, 0.35%. The use of this correction allowed weakly diffracting solvent to be modelled, with some success. Once the diffuse solvent correction had been applied, use of sharpened maps expedited the placing of further solvent sites in weak density. Without this correction, use of sharpened maps did not contribute to improvement of the model.

Low resolution fit of the model

Omission or incorrect modelling of solvent results in bad agreement between model structure factors and observed data at low resolution. The match of calculated and observed structure factors was assessed at sequential stages in the refinement of protein G, as listed in Table 5, as the treatment of solvent became progressively more sophisticated. Figures 4 *a*, *b, c* & *d* plot the fit of Fc to Fo for these models.

Table 5. Models used in assessment of fitting of the low resolution data

model	description
I	'unrefined' model from isotropic model 1igd (Derrick & Wigley, 1994) all solvent removed, random positional error of rms 0.3 Å applied to co-ordinates.
II	isotropic model refined from I using PROLSQ. Solvent network modelled using ARP.
III	anisotropic model refined without diffuse solvent correction, using SHELXL-93 and ARP.
IV	anisotropic model refined with the SHELXL-93 diffuse solvent correction applied, k_{sol} refined, B_{sol}, fixed.
V	anisotropic model refined with the ARP diffuse solvent correction applied; k_{sol} and B_{sol} both refined.

The plot of resolution dependence of (F_o/F_c), Figure 4*a*, shows an improvement of the fit at low resolution for ***II*** when compared with ***I***, due to the addition of discrete solvent during the isotropic refinement. F_c is much reduced, but still too large, after the isotropic refinement. The anisotropic model, without diffuse solvent correction, shows no improvement at low resolution, with respect to ***II***. However the models which have been refined with a diffuse solvent correction appear to be over-corrected, with $F_c < F_o$, in the low resolution range, for both ***IV*** & ***V***.

The R factor, Figure 4*b*, of the lowest resolution data is the same for the three refined models, ***II***, ***III*** & ***IV*** although, obviously, there is a progressive reduction of R at high resolution. The plot of $N\Delta^2$, Figures 4*c & d*, is most sensitive to differences in the quality of the models at low resolution. An improvement is seen for each level of sophistication in the solvent modelling, although the relative size of the improvements gets progressively smaller.

From the plot of (F_o/Fc) against resolution, it appears that both types of diffuse solvent correction are over enthusiastic. Changes in the solvent modelling do not strongly influence the low resolution R factor and $N\Delta^2$ is a much more sensitive indicator of improvements.

Distributions of solvent properties

The ordering of solvent in the rubredoxin and protein G structures described in Table 1 was compared. The rubredoxin model investigated contained 82 water molecules. If the volume unoccupied by protein were filled with pure water, the total solvent content of the unit cell would be 94 molecules. Therefore, virtually the entire solvent network may be modelled as a system of discrete sites and the amount of diffuse solvent must be minimal. The protein G solvent structure is more typical for a protein crystal. 119 water molecules were present in the model used in this comparison, while there is space for 222. Thus, almost half of the solvent is diffuse. This difference is reflected by the V_m

values for the structures. V_m for protein G, 2.23 Å^{3}Da^{-1}, lies in the middle of the range of the V_m distribution for protein structures, which is discussed above, while the rubredoxin value, 1.74 Å^{3}Da^{-1}, is at its lower limit.

The manner in which the crystal volume is divided into protein, ordered and diffuse solvent regions can be investigated by placing a grid inside the unit cell and ascertaining the distance of each grid point from the protein and from modelled solvent sites. Distributions of these distances were compared, Figures 5*a*, *b* & *c*. The histograms demonstrate the way in which the unit cell volume is apportioned between protein, hydrogen bonding shell around the protein and intermolecular space remote from the protein.

All the histograms, Figures 5*a*, *b* & *c*, have maxima in the 1.5 - 2.0 Å range, relating to a separation of 3-4 Å between atoms, which is the range of the van der Waal's radii for C, O and N atoms, while the area under the tail of the distribution, with $d > 2$ Å, gives information about characteristics of the solvent.

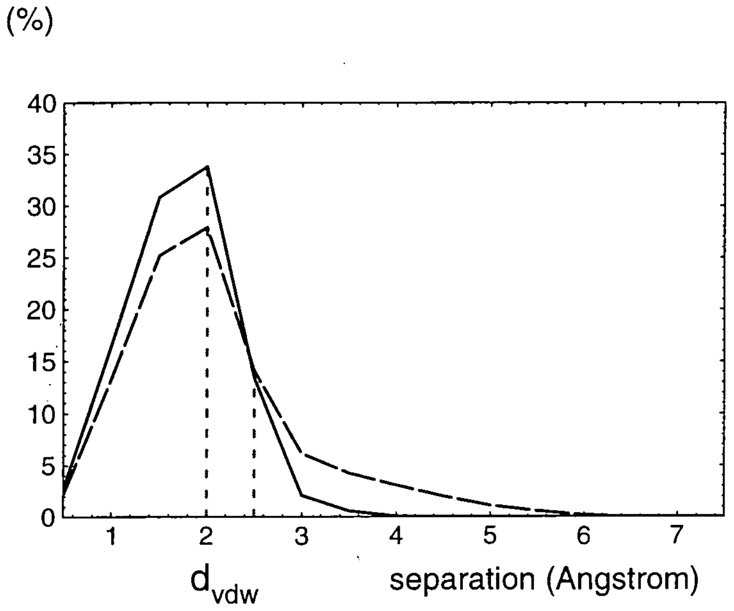


Figure 5.

An assessment of the division of the cell volume into protein, ordered and diffuse solvent regions, for rubredoxin and protein G models, Table 1. A grid was constructed inside the unit cell with divisions of around 0.5 Angstrom along each cell axis. A set of 128000 points was used. The rms separation between atom and closest gridpoint was around 0.25 Angstrom. Separations were computed for;

gridpoint to closest atom

gridpoint to closest protein atom

histograms were plotted of these lists of separations. The area under the histogram

for separations d1->d2 = unit cell volume separated by d1->d2 from all/protein atoms.

volume remote from the protein = volume of diffuse solvent region V(ds)

volume remote from all atoms = V(ra)

volume of the ordered solvent region, V(ss) = V(ds) - V(ra)

(a) Separation of gridpoints from closest atom, for rubredoxin (continuous line) and protein G (dashes).

Area under the plot with d > 2.5 Angstrom = V(ra)

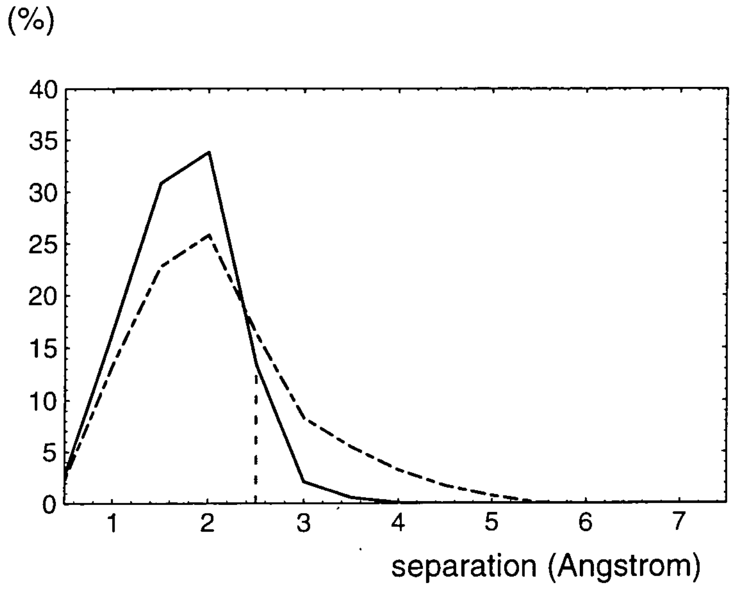


Figure 5b.

Separation of gridpoints from closest atom(continuous line) and protein atom(dashes), for rubredoxin. Area under continuous line plot with d > 2.5 Angstrom = V(ra), area between curves = V(ss)

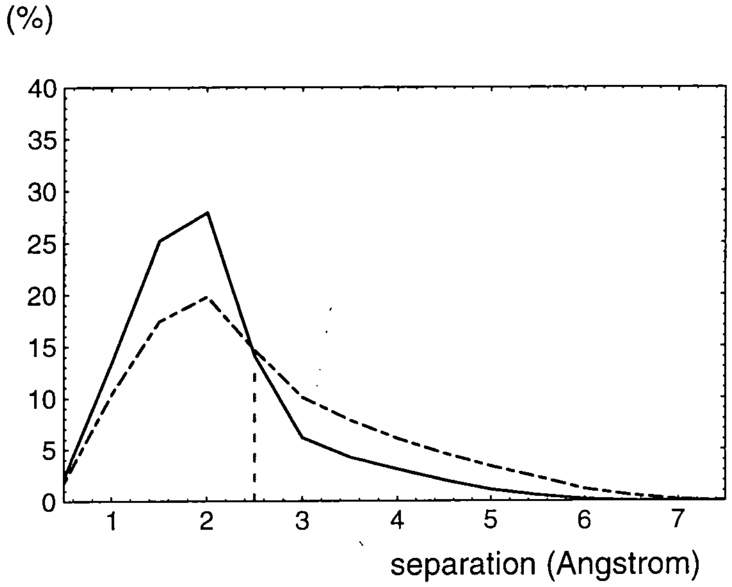


Figure 5c.

Separation of gridpoints from closest atom(continuous line) and protein atom(dashes), for protein G. Area under continuous line plot with d > 2.5 Angstrom = V(ra), area between curves = V(ss)

The results of this analysis of the solvent structure are summarised in Table 6. From the gridpoint analysis, the percentage of the cell volume occupied by the shell of ordered solvent, V_{ss}, is very similar for the two structures, 19.6% and 20.3%, for rubredoxin and protein G respectively. The percentage of ordered water in the cell can also be evaluated from V_{sc} and the total solvent content calculated from relative cell and protein volumes (9). $V_{sol,\ ordered}$ is 25 % and 24 % for rubredoxin and protein G. This gives a strong indication that the solvent shell around a protein has a standard conformation which is not greatly influenced by the nature of the protein or the packing density in the crystal, although the shell volume is obviously a function of the molecular surface area, which is relatively larger for a smaller protein.

$$V_{sol,\ ordered} = V_{sc}\ N_{sol,modelled} / N_{sol,total}\ (\%) \qquad (9)$$

From the gridpoint analysis, the percentage of the cell volume lying remote from the protein, V_{ds}, is 16.1% for rubredoxin and 31.4% for protein G, or, if V_{ds} is evaluated using gridpoint separation > 3 Å, 0.6% and 11.2% respectively. Using the alternative method (9), $V_{sol,\ diffuse}$ is 4% for rubredoxin and 21% for protein G. Thus, the protein G structure contains a large volume filled by diffuse solvent, while the rubredoxin structure has essentially no diffuse solvent region.

Table 6. Results of the analysis of solvent structure

	rubredoxin	protein G
from relative protein and cell volumes (Table 1):		
V_{sc} (%)	29	45
$N_{sol,modelled}$	82	119
$N_{sol,total}$	94	222
from gridpoint analysis (Figure 5)		
V_{ss} (%)	19.6	20.3
V_{ds} with d > 2.5 Å (%)	16.1	31.4
V_{ds} with d > 3 Å (%)	0.6	11.2
from V_{sc}, $N_{sol,modelled}$ & $N_{sol,total}$		
$V_{sol,\ ordered}$ (%)	25	24
$V_{sol,\ diffuse}$ (%)	4	21

A layered solvent structure, with solvent sites clustered at specific separation ranges from the protein, each cluster possessing a characteristic B factor range, will give rise to a series of peaks in the distribution of the function [B * (separation from protein)], Figure 6. For protein G, this distribution has two major peaks, at 100 and 150 Å^3, while the rubredoxin distribution has only one peak, at 100 Å^3. This suggests that a first solvent shell is present in both cases. Giving rise to the 100 Å^3 peak, it is situated in the hydrogen bonding region, at around 2.2 - 3.3 Å from the protein surface, with a characteristic B factor range, 25 - 35 Å^2. The second peak, present only in the protein G structure, could be caused by a second shell of water 3.3-4.0 Å from the protein surface with B factors in the region 35-50 Å^2.

The separation distance 3.3-4.0 Å is rather small for a second solvent shell and a value of 4.5 Å would be better. However solvent does indeed adopt a shell structure around the protein molecule in both crystals and rubredoxin is exceptional in having insufficient space between the molecules to incorporate more than a single shell.

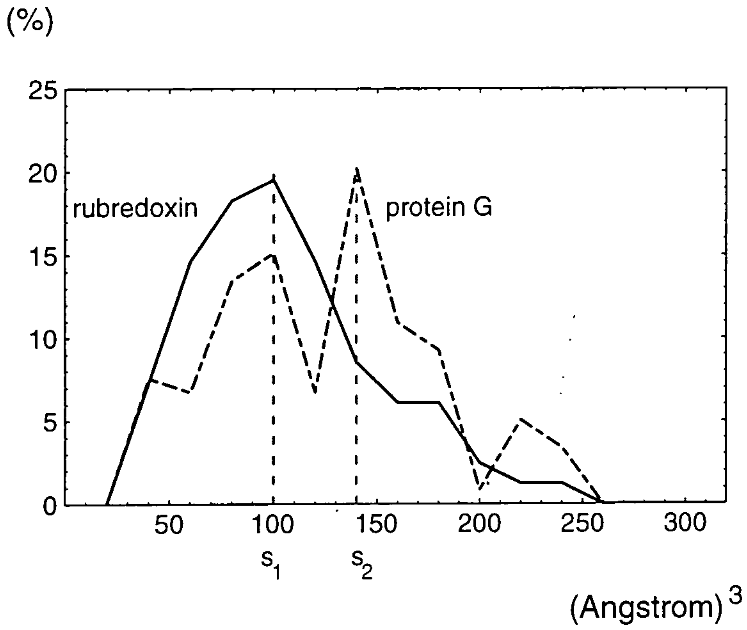


Figure 6.

The distribution of d(sp) * B(s) for rubredoxin(continuous line) and protein G(dashes), where d(sp) is the separation of a solvent site from its closest protein atom and B(s), the solvent site B factor. The existence of consecutive solvent shells, each possessing a characteristic B factor range, will give rise to a series of peaks in this distribution. Peaks are observed at separations s_1 and $s_2$for the protein G distribution, and s_1 only for rubredoxin.

Solvent Structure References

Dauter, Z., Sieker, L.C. & Wilson, K.S. (1992) Refinement of Rubredoxin from *Desulfovibrio vulgaris* at 1 Å with and without restraints. *Acta Crystallogr.* **B48**, 42-59.

Derrick, J.P. & Wigley, D.B. (1994) The third IgG-binding domain from streptococcal protein G. An analysis by X-ray crystallography of the structure alone and in a complex with Fab. *J. Mol. Biol.* **243**, 906-918.

Driessen, H., Haneef, M.I.J., Harris, G.W., Howlin, B., Khan, G. & Moss, D.S. (1989) RESTRAIN: restrained structure factor least-squares refinement program for macromolecular structures. *J. Appl. Cryst.* **22**, 510-516.

Karplus, P.A. & Faerman, C. (1994) Ordered water in macromolecular structure. *Current Opinion in Structural Biology* **4**, 770-776.

Lamzin, V.S. & Wilson, K.S. (1996) Automated refinement for protein crystallography. In *Methods Enzymol.: Macromolecular Crystallography.* (Carter, C.M. & Sweet, R.M. Eds.) in press.

Langridge, R., Marvin, D.A., Seeds, W.E., Wilson, H.R., Hooper, C.W., Wilkins, M.H.F. & Hamilton, L.D. (1960) The molecular configuration of Deoxyribonucleic Acid. II. Molecular models and their Fourier transforms. *J. Mol. Biol.* **2**, 38-64.

Levitt, M. & Park, B.H. (1993) Water: now you see it, now you don't. *Structure.* **1**, 223-226.

Matthews, B.W. (1968) Solvent content of protein crystals. *J. Mol. Biol.* **33**, 491-497.

Sheldrick, G.M. (1993) SHELXL-93, program for crystal structure refinement, *University of Göttingen,* Germany.

Tronrud, D.E. (1996) The TNT Refinement Package. In *Methods in Enzymology* (Carter, C. & Sweet, B. eds.) *in press.*

General References; Chapters 3, 4, & 5

Allen, F.H., Davies, J.E., Galloy, J.J.,Johnson, O., Kennard, O., Macrae, C.F., Mitchell, G.F., Smith, J.M. & Watson, D.G. (1991) *J. Chem. Inf. Comput. Sci.* **31**, 187-204

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Mayer, E.F., Bryce, M.D., Rodgers, J.R., Kennard, O., Simanouchi, T. & Tasumi, M. (1977) The Protein Databank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.

Blessing, R.H. & Langs, D.A. (1988) *A priori* estimation of scale and overall anisotropic temperature factors from the Patterson origin peak. *Acta Crystallogr.* **A44**, 729-735.

Blessing, R.H. & Langs, D.A. (1996) Statistical expectation value of the Debye-Waller factor and E(*hkl*) values for macromolecular crystals. *Acta Crystallogr.* **D52**, 257-266.

Blundell, T.L., & Johnson, L.N. (1976) Protein Crystallography. *Academic Press, London.*

Brünger, A.T. (1990) X-plor manual, version 2.1. *Yale University.*

Brünger, A.T. (1993) Assessment of phase accuracy by cross validation: the free R value. Methods and application. *Acta Crystallogr.* **D49**, 24-36.

CCP4 (1979) The SERC (UK) Collaborative Computational Project Number 4, a suite: programs for protein crystallography. Daresbury laboratory, Warrington, UK.

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

Cochran, W. (1952) A relation between the signs of structure factors. *Acta Crystallogr.* **5**, 65-67.

Dauter, Z., Lamzin, V.S. & Wilson, K.S. (1995) Proteins at atomic resolution. *Current Opinion in Structural Biology* **5**, 784-790.

Dodson, E., Kleywegt, G. J. & Wilson, K. (1996) Report of a workshop on the use of statistical validators in protein crystallography. *Acta Crystallogr.* **D52**, 228-234.

Driessen, H., Haneef, M.I.J., Harris, G.W., Howlin, B., Khan, G. & Moss, D.S. (1989) RESTRAIN: restrained least-squares refinement program for macromolecular structures. *J. Appl. Cryst.* **22**, 510-516.

Engh, R.A. & Huber, R. (1991) Accurate bond and angle parameters for X-ray protein structure refinement. *Acta Crystallogr.* **A47**, 392-400.

Hendrickson, W.A. (1985) Stereochemically restrained refinement of macromolecular structures. *Meth. Enzymol.* **115**, 252-270.

James, R.W., (1948) False detail in three-dimensional Fourier representations of crystal structures. *Acta Crystallogr.* **1**, 132-134.

Jones, A.T. (1978) A graphics model building and refinement system for macromolecules. *J. Appl. Crystallogr.* **11**, 268-272.

Jones, T.A., Zou, J.Y., Cowan, S.W & Kjeldgaard, M. (1991) Improved methods for building models in electron density maps and the location of errors in these models. *Acta Crystallogr.* **A47**, 110-119.

Karle, J. & Hauptmann, H. (1953) Application of statistical methods to the naphtalene structure. *Acta Crystallogr.* **6**, 473-476.

Karle, J. & Hauptmann, H. (1956) A theory of phase determination for the four types of non-centrosymmetric space groups *1P222, 2P22,* $3P2_12$, $3P_22$. *Acta Crystallogr.* **9**, 635-651.

Konnert, J.H. & Hendrickson, W.A. (1980) A restrained-parameter thermal-factor refinement procedure. *Acta Crystallogr.* **A36**, 344-350.

Kraulis, P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* **24**, 946-950.

Lamzin, V.S. & Wilson, K.S. (1993) Automated refinement of protein models. *Acta Crystallogr.* **D49**, 129-147.

Lamzin, V.S., Sevcik, J., Dauter, Z. & Wilson, K.S. (1995) Implications of atomic resolution. *Making the most of your model. Proceedings of the CCP4 Study Weekend, 6-7 January, SERC Daresbury Laboratory, Daresbury, Warrington, England,* 33-40.

Lamzin, V.S. & Wilson, K.S. (1996) Automated refinement for protein crystallography. In *Methods Enzymol: Macromolecular Crystallography.* (Carter, C.M. & Sweet, R.M. Eds.) in press.

Langridge, R., Marvin, D.A., Seeds, W.E., Wilson, H.R., Hooper, C.W., Wilkins, M.H.F. & Hamilton, L.D. (1960) The molecular configuration of Deoxyribonucleic Acid. II. Molecular models and their Fourier transforms. *J. Mol. Biol.* **2**, 38-64.

Laskowski, R.A., MacArthur, M.W., Moss, D.S. & Thornton, J.M. (1993) PROCHECK. *J. Appl. Cryst.* **26**. 283-291.

Lunin, V.Y. (1993) Electron-density histograms and the phase problem. *Acta Crystallogr.* **D49**, 90-99.

Main, P. (1990) A formula for electron density histograms for equal-atom structures. *Acta Crystallogr.* **A46**, 507-509.

Matthews, B.W: (1968) Solvent content of protein crystals. *J. Mol. Biol.* **33**: 491-497.

Murshudov, G., Dodson, E. & Vagin, A. (1996). Program and examples on maximum likelihood refinement. *Proceedings of the CCP4 Study Weekend (Bailey, S., Dodson, E.,& Moore, M., eds.) SERC Daresbury Laboratory, Warrington, UK.*

Navaza, J. (1994) AMoRe: An automated package for molecular replacement. *Acta Crystallogr.* **A50**, 157-163.

Otwinowski, Z. & Minor, W. (1993) DENZO: a film processing program for macromolecular crystallography. *Yale Univ., New Haven, CO., USA.*

Podjarny, A.D. & Yonath, A. (1977) Use of matrix direct methods for low-resolution phase extension for tRNA. *Acta Crystallogr.* **A33**, 655-661.

Richardson, J.S. (1981) The anatomy and taxonomy of protein structure. *Advan. Protein Chem.* **34**, 167-339.

Rollett, J.S. (1970) Least-squares procedures in crystal structure analysis. Crystallographic computing, *Munksgaard, Copenhagen.*

Sheldrick, G.M., Dauter, Z., Wilson, K.S., Hope, H. & Sieker, L.C. (1993) The application of direct methods and Patterson interpretation to high-resolution protein data. *Acta Crystallogr.* **D49**, 18-23.

Sheldrick, G.M. (1993) SHELXL-93, program for crystal structure refinement, *University of Göttingen,* Germany.

Sheldrick, G.M. & Schneider, T.R. (1996) SHELXL: High resolution refinement. *Methods Enzymol. , in press.*

Stenkamp, R.E. & Jensen, L.H., (1984) Resolution revisited: limit of detail in electron density maps. *Acta Crystallogr.* **A40**, 251-254.

Stout, G.H. & Jensen, L.J. (1989) X-ray structure determination: a practical guide. 2nd edition. *Wiley, New York.*

Stryer, L. (1988) *Biochemistry, 3rd edition.* W.H. Freeman & Co. New York.

Swanson, S.M., (1988) Effective resolution of macromolecular X-ray diffraction data. *Acta Crystallogr.* **A44**, 437-442.

Ten Eyck, L. (1973) Crystallographic fast Fourier transforms. *Acta Crystallogr.* **A29**, 183-191

Tronrud, D.E. (1996) The TNT Refinement Package. In *Methods in Enzymology* (Carter, C. & Sweet, B. eds.) *in press.*

Vriend, G. & Sander, C. (1993) Quality control of protein models: directional atomic contact analysis. *J. Appl. Cryst.* **26**, 47-60.

Wilson, A.J.C. (1942) Determination of absolute from relative X-ray data intensities. *Nature* **150**, 151-152.