Studies on the Crystallographic Phasing of Proteins: Substructure Validation and MAD-phased Electron Density Maps at Atomic Resolution

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten
der Georg-August-Universität zu Göttingen

vorgelegt von
Fabio Dall'Antonia
aus Göttingen

Göttingen 2003

D7

Referent: Prof. Sheldrick Ph. D.

Korreferent: Prof. Dr. Ficner

Tag der mündlichen Prüfung: 06.11.2003

Contents

1 Introduction 1

1.1 Macromolecular crystal structure determination at high resolution 1
1.2 Aldose Reductase 3
1.3 The solution of the phase problem with experimental methods 6
1.4 Motivation and scope of the thesis 7

2 Theoretical Background 9

2.1 Experimental methods to solve the Crystallographic Phase Problem 9
2.1.1 X-ray diffraction on crystals and the structure factor 9
2.1.2 The experimental phasing of macromolecules with heavy-atom derivatives 11
2.1.3 Practical aspects of experimental macromolecule phasing 22
2.2 Macromolecular structure refinement 27

3 Materials and Methods 31

3.1 Studies on the crystal structure of human Aldose Reductase 31
3.1.1 Expression, purification and crystallization of Aldose Reductase 31
3.1.2 Data collection and reduction 31
3.1.3 Data analysis and exploitation 31
3.1.4 The localization of selenium sites 33
3.1.5 Heavy atom model refinement and protein phase calculation 34
3.1.6 Density modification by solvent flattening 37
3.1.7 Refinement of the hAR protein structure with SHELXL 38
3.1.8 Comparisons of phases and electron density maps 46
3.1.9 The creation of an experimentally phased difference electron density map 47
3.2 The development of the substructure validation program SITCOM 48
3.2.1 Definition of SITCOM 48
3.2.2 Program architecture 48
3.2.3 Program flow 48
3.2.4 General algorithms for site comparison 51
3.2.5 Polar spacegroups 52
3.2.6 The P1 algorithm 53
3.2.7 Substructure scoring 54

4 Results and Discussion 59

4.1 Results for the crystal structure of human Aldose Reductase 59
4.1.1 Refinement characteristics and description of the refined hAR2 model 59
4.1.2 Evaluation of the experimental phases derived from hAR data 81
4.1.3 Analysis of the experimental electron density maps 85
4.1.4 Classification of disorder using the experimental map 90
4.2 The Application of SITCOM 101
4.2.1 The verification of program functionality with test structures 101
4.2.2 Studies on Transhydrogenase B 101
4.2.3 SAD phasing of HAPTBr 107

5 Summary and Conclusions 121

5.1 Aldose Reductase 121
5.2 SitCom 122
5.3 Future perspectives and final remarks 124

6 References 127

Appendices 133

List of Abbreviations

APS	*Advanced Photon Source* (Argonne, Illinois, U.S.A)
CC	correlation coefficient
C-terminal	carboxy-terminal
E. coli	*Escherichia coli*
et al.	*et alii*
F_c	calculated structure factor
F_o	observed structure factor
FFT	fast fourier transform
Fig.	figure
FOM	figure of merit
HAPTBr	*Human Acyl Protein Thioesterase I*
hAR2	*human Aldose Reductase*
IDD	*Institute for Diabetes Discovery* (Branford, CT, U.S.A)
kDa	kilo-Dalton
LL	log-likelihood
LS	least squares
MAD	Multiple Anomalous Dispersion
MIR	Multiple Isomorphous Replacement
NCS	non-crystallographic symmetry
N-terminal	amino-terminal
PatFOM	Patterson-figure of merit
PDB	Protein Data Base
r.m.s	root mean square
SAD	Single Anomalous Dispersion
SIR	Single Isomorphous Replacement
Tab.	table
THB	*Transhydrogenase B*

Software used for structure visualization

Figures containing molecular graphics were created with the program Raster3D (Merrit & Bacon 1997) after preparation with the programs BobScript (Esnouf 1997), DINO (A. Philippsen, http://www.dino3d.org), or Molscript (Kraulis 1991).

Figs. 4.43, 4.48 and 4.49 were created with Rasmol (Sayle & Milner-White 1995).

Figs. 4.10, 4.36, 4.38 and 4.41 were created with Ortep3 (L. Farrugia, http://www.chem.gla.ac.uk/~louis/ortep3/) and POV-Ray (Persistence of Vision, http://www.povray.org).

Fig. 4.13 was created with LigPlot (Wallace et *al.* 1995).

1 Introduction

The present thesis is concerned with methodological studies related to the experimental determination of crystallographic macromolecule phases. While one part of the work has been the development and application of a validation program for heavy atom substructure solutions, the properties of highly resolved, experimentally phased electron density maps have been investigated in the other. In the following, a brief introduction to the current state and limits of macromolecular crystallography and a short overview on experimental phasing are leading to the scope and motivation of the thesis.

1.1 Macromolecular crystal structure determination at high resolution

With the recent achievements in genome sequencing and the challenges for medically oriented biochemistry research, structural biology and in particular biomolecular crystallography have obtained a interdisciplinary key role for the availability of new functional insights on a molecular level. Thanks to modern genomics, the number of sequenced, yet functionally and structurally unknown proteins is still rising considerably. On the other hand, the comprehension of substrate-enzyme and inhibitor-enzyme interactions on a structural basis has been made essential by the need for drug targets.

The method of X-ray crystallography is superior to other structure determination methods with respect to the information content and accuracy obtained as well as the range of applicability. Electron microscopy, for example, is restricted to a resolution limit of about 8 Å, far away from an atomar level. Nuclear magnetic resonance (*NMR*) is a high-resolution method with an accuracy comparable to X-ray crystallography, but it has the decisive drawback of being limited to structures of less than 30kDa molecular weight. On the contrary, X-ray crystallography has tackled ever larger structural problems – an outstanding example has been the determination of the two ribosomal subunits (Wimberly et *al.* 2000, Ban et *al.* 2000). The only major restriction of method is given by its name – it is the dependence on the availability and quality of crystals.

The key requirement of crystallographic X-ray structure determination is the interpretablility of electron density obtained from the diffraction experiment. The most important factor for the accuracy and information content of a crystal structure is the resolution of the electron density map.

If the structural resolution is very low, i.e. below 6 Å, only the location of large domains and the rough shape of the tertiary protein structure can be identified. This resolution area is becoming a border case, where high-performance electron microscopy can contribute to observations. In the low resolution range between 3 and 4 Å, secondary structure elements become visible, and α-helices can be distinguished from β-sheets. This level of recognition already allows the identification of the protein fold and possibly some predictions about functionality, because tertiary structure motifs can be assigned to protein families of known functional properties. Beyond a resolution of 3 Å, individual amino acids can be identified and below 2 Å, even single atom positions become visible. The structural interpretability develops from a level of biological relevance to the availability of chemically relevant information, such as atom type recognition, bond length determination and identification of non-covalent interactions, in particular hydrogen bonds. All these structural properties are needed to describe detailed biochemical functions, for example the reaction mechanism of an enzyme. From the crystallographic point of view, the medium-to-high resolution region is interesting, because individual parameter refinements for all atoms are possible. Moreover, crystal-specific features like positional disorder become observable. Reflecting conformational flexibility, the presence or absence of disorder can answer questions about floppy or fixed protein regions related for example to substrate interactions.

Which features, requiring even higher resolution, remain? The atomic resolution area around and beyond 1 Å reveals structural details that are normally confined to small molecule crystallography – for example, the large number of structure factor data enables the refinement of anisotropic displacement parameters (Dauter et al. 1997). The directions of atomic displacements are important, because they may indicate fine nuances of disorder, in particular positional protein backbone shifts, that could be connected to substrate binding.

Covalent hydrogen bonds are shorter than 1 Å. Therefore, electron density maps of (sub-) atomic resolution theoretically reveal the positions of hydrogen atoms. The modeling of geometrically ideal hydrogen atoms could be abandoned in favour of atom placement according to difference density peaks. Experimentally confirmed hydrogen atom positions are of extreme interest, if their presence (or absence) proves reaction mechanisms. This situation occurs for the 0.9 Å structure of *Aldose Reductase*, as discussed later. It has to be emphasized however, that even given a sufficient resolution, the hydrogen atom localization is problematic because of the small atomic scattering contribution of only one electron per atom.

In principle, geometrical restraints are not needed if the resolution provides a sufficiently high data-to-parameter ratio. It will be shown that the structure refinement of *Aldose Reductase* revealed examples, where the observed and refined sidechain geometry violated restraints, but could be confirmed with experimentally phased electron density maps. This domination of “real” data over restraints, used as data substitutes, does not only question the justification of (too strict) restraints at atomic resolution – of course it also gives rise to the question whether the values of restraints, often derived from small molecule structures – are necesserily appropriate for protein structures, too.

There are two major requirements for the collection of highly resolved diffraction data. The first is high quality of the crystal itself, *i.e.* a well ordered crystal lattice with low mosaicity, the second is high quality of the X-ray instrumentation. This is based on the fact that the diffraction event itself is determining the quality of the diffracted reflections, and the two components interacting during the scattering process are the X-ray beam and the crystal medium.

Experience has shown that the crystal quality can be influenced or adjusted only to a rather small extent. The prediction of crystal properties and diffraction behaviour from crystallization conditions is often very difficult. More progress has been made in the development of high-quality X-ray instrumentation. In particular, the intensity and brillance of the X-ray beam has improved considerably with the availability of synchrotron radiation from third generation facilities during the last few years. Apart from this, the sensitiveness and precision of X-ray area detectors has become better and the use of cryo systems, reducing the thermal motion of the atoms, provides a more accurate structure determation as well.

1.2 Aldose Reductase

The Aldo-Keto Reductase superfamily consists of enzymes of Mw ~ 36 kDa which catalyze the reduction of various substrates containing an aldehyde functionality. Among the aldo-keto reductases, *Aldose Reductase* (*AR2*) is one of the most thoroughly studied proteins.

Human Aldose Reductase (*hAR2*) catalyzes the metabolic reduction of glucose to sorbitol that takes place as a hydrogen transfer from NADPH+H$^+$. The subsequent oxidation of sorbitol to fructose does not occur in cell tissues lacking the corresponding enzyme *Sorbitol Dehydrogenase*. Thus, a high glucose concentration in nerve or eye lens tissues, as caused by *Diabetes Mellitus,* leads to an excess of sorbitol and diabetic symptoms such as neuropathy or cataract result (Gonen & Dvornik 1995). Pharmacological studies have shown that hAR inhibitors significantly reduce the enzyme activity and may be applied to prevent diabetic complications (Dvornik 1994).

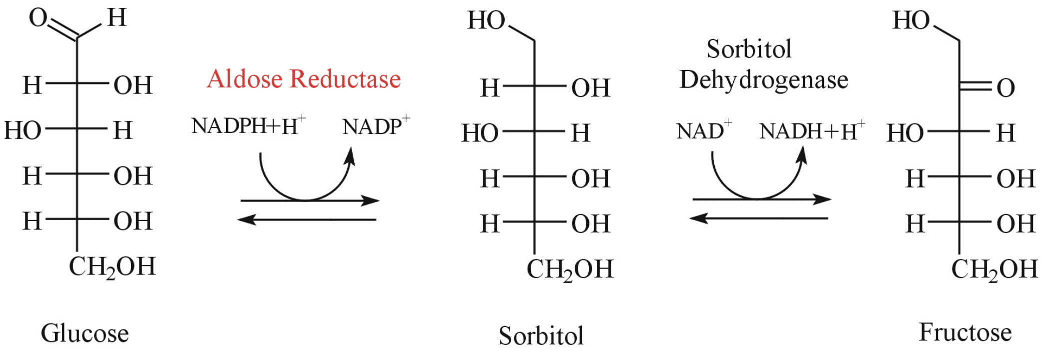


Fig. 1.1: The sequence of chemical reactions responsible for the transformtion of glucose to fructose. The rever sible reactions are catalized by the enzymes Aldose Reductase and Sorbitol Dehydogenase

Aldose Reductase is an enzyme consisting of 316 amino acid residues and exhibiting the $(\beta\alpha)_8$ TIM barrel fold common to all members of the family (see chapter 4.1.1.1).

Crystallography on Aldose reductase has been focussed on the active site features relevant to drug design. Several inhibitor complexes have been studied, some of which lead to crystal structures (e.g. Urzhumtsev et *al*. 1997, Calderone et *al*. 2000, El-Kabbani et *al*. 2003). It has been shown that the residues Tyr48, His110 and Trp111 are involved in the inhibitor binding and that at least ten other residues indirectly contribute to the shape of the active site. It was also found by site-directed mutagenesis (Bohren et *al.* 1994) and by modelling studies (Lee et *al.* 1998) that His110 and Tyr48 are the two possible proton donors for the substrate – however, clear statements about a preference for either of the two residues could not be made.

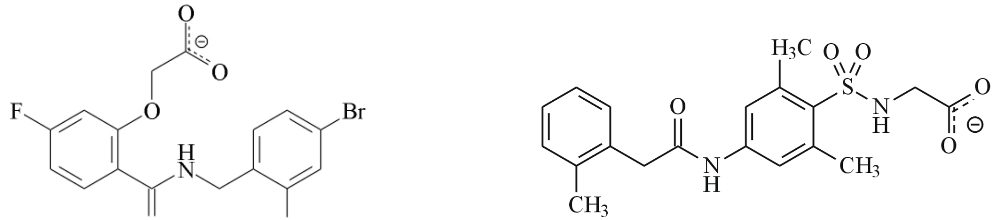


Fig.1.2: Two hAR2 inhibitors with a carboxylic "head" as common feature, *IDD594* (left) and *IDD384* (right)

Due to its negatively charged carboxylic "head", the inhibitor molecule *IDD594* establishes particularly strong hydrogen bonds to the three residues Tyr48, His110 and Trp111, mentioned before. Crystals of ternary *hAR2-NADP*$^+$*-IDD594* complexes obtained at pH 5.0 have lead to the best crystallographic *hAR2* results so far, including the 0.9 Å Seleno-Met derivative MAD structure studied in the present work, a 0.8 Å Seleno-MAD structure determined from several different crystals and a 0.62 Å native structure, refined to 0.66 Å (Howard et *al.* submitted to *Proteins*), representing the highest resolution ever of a crystal structure of 36 kDa size. The extremely good scattering power of hAR2 crystals is obviously favoured by their considerable size ($0.3*0.4*0.6$ mm^3) and may additionally be explained by intrinsic features of the structure, for example a low solvent content of about 30% and a loop region which is positionally fixed because of the inhibitor interactions. During data collection, the use of high-performance X-ray instrumentation, such as an ondulator beamline at the 3rd generation APS synchrotron, contributed as well to diffraction beyond atomic resolution.

Based on the 0.66 Å refinement model and a combination of molecular dynamics and quantum mechanic calculations, a catalytic proton transfer mechanism (Fig. 1.3) explaining an active role of both Tyr48 and His110 was proposed (Cachau et *al*. 2000). The mechanism model, although agreeing well with available crystallographic data, still lacks a direct experimental proof of the hydrogen transfer steps. One problem in this context is hitherto the unavailability of a sufficiently high-resolved crystal structure at the physiological pH 7.0, which could confirm the exact position and protonation state of the key residue His110. Another retention against the proposed mechanism is the generally problematic hydrogen identification in crystallography, even at sub-atomic resolution, as mentioned previously. All relevant hydrogen atoms have been observed in electron density omit maps from the 0.66 Å refinement, but it has to be pointed out that the use of such maps is prone to model bias. This is because the calculated structure factor phases used for omit maps tend to maintain contributions of removed hydrogen atoms. The crystallographic proof of the hAR2 reduction mechanism would become more reliable, if bias-free data was used. Therefore, the interpretation of a high-quality experimental (MAD) map of atomic resolution might turn out to be highly valuable for the confirmation of the hydrogen atom model.

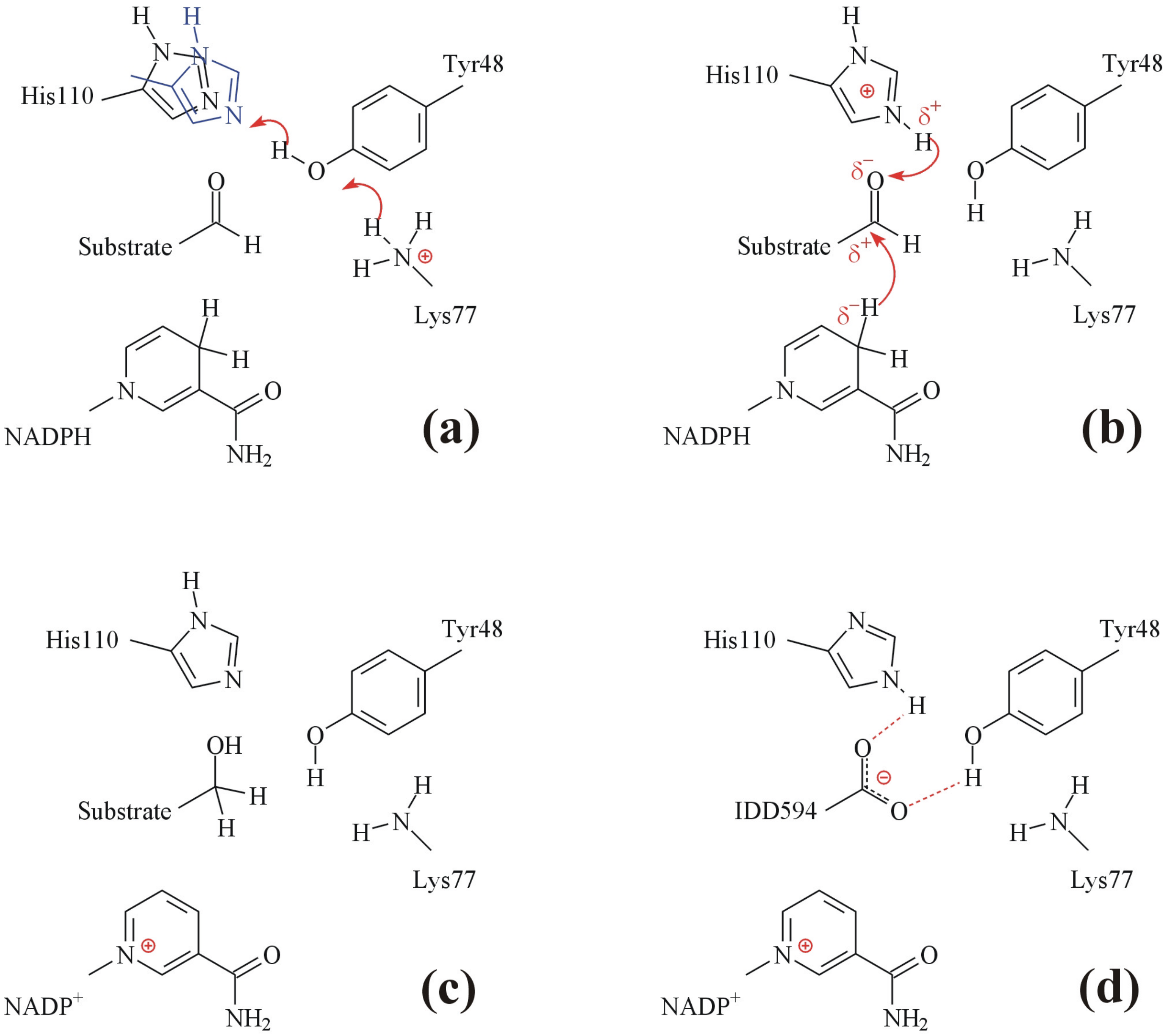


Fig.1.3: The mechanism of substrate reduction, catalized by hAR2, as proposed by Cachau et al. (a) Tyr48 acts as initial proton source, transfering its H$^+$ to His110. The model calculations have yielded that this key step requires the tilt of His110 from its usual position, found in the crystal structure, towards Tyr48. The protonated Lys77 is donating its H$^+$ imidiately to Tyr48, which is unchanged after the step. (b) In the subsequent step, the substrate is reduced by the nicotinamide moiety of NADPH acting as hydride donor. Upon reduction of the electrophilic aldehyde carbon atom, the nucleophilic carbonyl oxygen atom is protonated by His110. (c) In the resulting state, the former substrate aldehyde (Glucose) has become an alcohol (Sorbitol). NADPH has lost a H$^-$ to become NADP$^+$. Lys77 has lost a proton – to complete the catalysis cycle, Lys77 has to be re-protonated. (d) The protonation states found in the hAR2-IDD594 crystal structure. The negatively charged carboxylic head of the inhibitor makes hydrogen bridge contacts to His110 and Tyr48, thus blocking the protonation site. The inhibitor itself is not reducable, and therefore non-competitive.

1.3 The solution of the phase problem with experimental methods

Experimental phasing methods comprise the different techniques of using macromolecular heavy atom derivatives for structure solution – either as *isormorphous replacement* (*SIR, MIR*), or by exploiting the *anomalous dispersion* of heavy atoms (*SAD, MAD*), or by a combination thereof (*SIRAS, MIRAS*). It is common to all these methods, that the protein phase estimates are derived from the experimental observables, namely measured X-ray reflection intensities, without external phase information. Thus, they can be distinguished from statistical *ab initio* methods and from *molecular replacement* methods.

Some of the experimental phasing methods, SIR and SAD, only yield ambiguous phase estimates of low reliability. Especially for these methods, the subsequent phase improvement by density modification techniques (e.g. Wang 1985) is essential. In general, the quality of derived phases critically depends on the accuracy and precision of the intensity measurements. The progress in the experimental data collection strategies, with a focus on high redundancy of the measured intensity data, has therefore contributed to the applicability of SAD as a successful structure solution method (Debreczeni et *al.* 2003). There have been improvements in the use of derivatives from heavy atom soaks, in particular halide soaks (Dauter & Dauter 1999) as well as approaches exploiting the anomalous scattering contributions of atoms in native crystals, namely sulfur in proteins (Dauter et *al.* 1999, Yang & Pflugrath 2001) and phosphorus in DNA (Dauter & Adamiak 2001).

The *multiple-wavelength anomalous dispersion* (MAD) method, compared to SAD, has the great advantage of yielding unambiguous estimates of the structure factor amplitudes for the heavy atom substructures (F_A, Hendrickson et al. 1985), from which much more reliable protein phases can be derived. The MAD phasing method has become a routine procedure which allowed to solve increasingly large structures (e.g. van Delft 2003).

For MAD diffraction data of good quality, it was found that the use of anomalous difference estimates (ΔF) obtained from single-wavelength data subsets ("pseudo-SAD" data) alone can provide phases of a quality comparable to F_A data . Taking advantage of the calculation speed of modern computer hardware, structure solution from the first collected MAD data subset can be attempted while the experiment is still performed. (Dauter 2002). No matter if F_A or ΔF estimates are used, the determination of the substructure of anomalous scatterers is a crucial step. The completeness and the accuracy of the found heavy atom sites significantly influence the final protein phase reliability. In this context, the comparative analysis of heavy atom substructures determined from the F_As and the ΔFs of individual data collections at different MAD wavelengths would be valuable to establish future strategies for SAD and MAD phasing.

Usually, the resolution limits of reflection data obtained from derivative crystals are lower than for native crystals. Therefore, the traditional structure determination strategy based on experimental methods consists of heavy atom substructure solution and initial phase calculation from a derivative dataset, followed by phase improvement, phase extension, and structure refinement against native diffraction data of higher resolution. The possibilities provided by modern synchrotron X-radiation (increased beam brillance and intensity as well as a tuneable wavelength), allow the MAD data collection to resolution limits suited for the named tasks. The structure factor amplitudes taken from the most reliable MAD data subset (usually measured at high-energy remote wavelength) can be used for structure refinement, making native data in principle superfluous. Much greater advantage however can be taken from the experimentally determined phases, if they are of high reliability and extend to high resolution. In such cases, electron density maps without any influence of prior model information can be calculated and used to investigate structural features.

The observation of solvent shells and multiple protein conformations in MAD-phased electron density maps resolved to 1.8 Å (Burling et *al.* 1996) and even 1.0 Å (Schmidt et *al.* 2002) have been reported as well as the use of 1.2 Å SAD phases (extended to 1.0 Å, Thaimattam et *al.*, to be published) and 1.1 Å SAD phases (Brodersen et *al.* 2000) to obtain model-independent experimental maps. These advances in resolution of experimental phases exemplify the value of the method for the direct interpretation of experimental electron density or the validation of previously modelled structural features.

1.4 Motivation and scope of the thesis

Taking the importance of accurately determined heavy atom substructures for protein structure solution with the *SAD* or *MAD* method into account, a substructure comparing computer program called SITCOM was developed, aiming at the improved applicability of substructure solutions from various programs, e.g. SnB (Weeks & Miller 1999), SHELXD (Schneider & Sheldrick 2002) or SOLVE (Terwilliger & Berendzen 1999): In case of problematic crystal structures providing weak diffraction data, situations may occur where a straight-forward substructure determination becomes tedious or even impossible. Often, this is a problem of identifying a correct solution in a pool of many trials, or a problem of interpreting a given substructure correctly (*i.e.* of selecting the correct heavy atom sites), rather than a problem of solving the structure with the respective program. To help the "manual" identification process, SITCOM automizes the comparison of substructure solution trials from one or several progams in order to select the best solution and to pick the most reliable sites from that solution.

Even between similar (roughly correct) solutions, the accuracy of heavy atom sites may vary, thus influencing the phasing results. An absolute determination of site accuracy in terms of analysis of distances to the true heavy atom positions is only possible, if a solved and refined protein structure exists. The *a-posteriori* comparison between refined and experimental heavy atom site positions from different reflection data sources, respectively heavy atom structure factor estimates ($MAD\text{-}F_A$ or $SAD\text{-}\Delta F$ from different wavelengths), is very valuable to jugde the corresponding substructure accuracy and develop recommendations for the experimental structure solution strategy.

Therefore, in the studies of the present thesis, SITCOM was also applied to compare data subset- and resolution-dependent Selenium substructures from SHELXD to the refined Selenium positions for the proteins *Transhydrogenase B* (Buckley et *al.* 2000) and human *Acyl-Protein Thioesterase I* (Devedjiev et *al*. 2000).

As mentioned before, in case of protein derivative diffraction data of very high resolution and quality, the value of experimentally derived phases exceeds the single use of an initial electron density map for model building. The Seleno-Methionine derivative structure of human Aldose Reductase in complex with inhibitor IDD594, as obtained from the 0.9 Å MAD data, was refined against structure factor amplitudes from the high-energy remote data subset. Using the model-independent experimental map calculated from the MAD phases, the modelled features were verified in detail afterwards, particularly focussing on the validation and classification of disorder.

The present thesis, with its two parts, is thus investigating some possibilities of method improvement for experimental phasing are as well as the possible benefits of such improved methods for structural biology. In the following two major chapters, both divided in sections about SITCOM and Aldose Reductase, first the used materials and methods are explained, and afterwards the obtained results are presented and discussed.

2 Theoretical Background

2.1 Experimental methods to solve the Crystallographic Phase Problem

2.1.1 X-ray diffraction on crystals and the structure factor

The phenomenon of X-ray diffraction is based on the interaction between electromagnetic radiation and the electron shells of atoms. Looking at a single atom, interfering X-ray quantums will excite harmonic oscillation of electrons causing the simultaneous emission of scattered waves. The propagation of these waves is spherical, *i.e.* into all space directions. The wavelength of the scattered beam remains the same as for the original X-ray beam. The scattering depends on the distribution of electron density ρ around the atom core, which is a function of the atomic radius r. The scattering is also dependent on the angle of incidence Θ and the wavelength λ of the X-rays. The integration over the electron shell of an atom (with the simplifying assumption of a spherical shape) leads to the the atomic scattering factor f_a:

$$f_a = 4\pi \int_0^\infty \rho(r)\; r^2 \frac{\sin\left(2\pi r \frac{2\sin\Theta}{\lambda}\right)}{2\pi r \frac{2\sin\Theta}{\lambda}} dr \qquad\qquad f_a' = f_a \exp\left(-\frac{B\sin^2\Theta}{\lambda^2}\right)$$

The distribution of electron density becomes more complex, if the thermal motion of the atom is taken into account. Therefore, the corrected scattering factor f_a' is introduced, containing the exponential *Debye-Waller* term. The temperatur factor B used in macromolecular crystallography is connected to the squared displacement parameter u^2 via $B = 8\pi^2u^2$.

X-ray scattering leads to X-ray diffraction, if many atoms are arranged in a three-dimensional lattice. This is the case for the crystalline state of matter. If X-rays hit a crystal, scattered beams may interfere constructively and give rise to a so-called reflection. Certain orientations of net planes (sections of the crystal lattice) relative to the beam cause different reflections, therefore diffraction patterns are observed upon the detection of scattered X-rays. The *Bragg* equation defines the X-ray incidence angles to the lattice that actually lead to diffraction. Each reflection is decribed by a structure factor F_{hkl}, where the *Miller indices* h, k and l represent the orientation of the net planes as fractions of crystal unit cell edges. Diffraction is caused by electrons, therefore the structure factor can be understood as the integral over the electron density function ρ(x,y,z) in the unit cell:

$$F_{hkl} = \int_{cell} \rho(xyz)\exp[2\pi i(hx + ky + lz)]\; dV$$

The electron density is not distributed arbitrarily, but connected to the location of atoms. Thus, discrete maxima of electron density at the atomic positions can be assumed. Alternatively to an integration, the structure factor is described by a fourier transform of the atomic electron density, summing over all individual atomic scattering factors:

$$F_{hkl} = \sum_{i=1}^{N} f'_i \left\{\cos[2\pi(hx_i + ky_i + lz_i)] + i\sin[2\pi(hx_i + ky_i + lz_i)]\right\}$$

Structure factors are complex quantities, consisting of a real and an imaginary component. This is best elucidated by representing the structure factor as a vector in the complex plane:

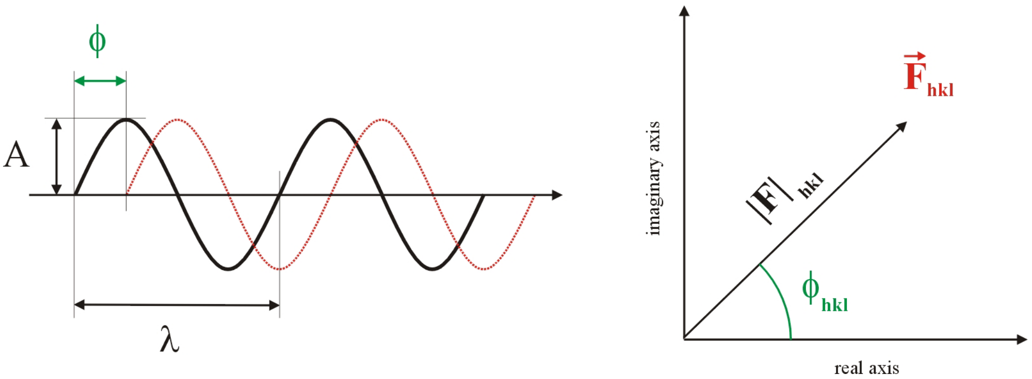


Fig. 2.1: Representation of a reflection as wave (left) and complex structure factor vector (right)

As a reflection is consisting of scattered X-ray waves, it can also be represented in form of a wave, defined by amplitude A, wavelength λ (i.e. frequency) and phase ϕ. The amplitude of a reflection is equal to the the structure factor vector length, $|F_{hkl}|$, while the phase shift is given by the phase angle ϕ_{hkl} in the complex plane. Separating the real and the imaginary component of the structure factor, i.e. amplitude and phase, another description of the structure factor is obtained:

$$\vec{F}_{hkl} = |F_{hkl}|\exp(i\Phi_{hkl}) \qquad \Phi_{hkl} = \arctan\left(\frac{\sum_i f_i'\sin[2\pi(hx_i + ky_i + lz_i)]}{\sum_i f_i'\cos[2\pi(hx_i + ky_i + lz_i)]}\right)$$

The phase shift is containing the information about the atomic positions. Thus, to calculate the electron density function at any given location x, y, z of the unit cell from the structure factors, a reverse fourier transform has to be made. The summation over all structure factors F_{hkl} requires both amplitudes and phases:

$$\rho(xyz) = \frac{1}{V}\sum_{hkl}|F_{hkl}|\exp[-2\pi i(hx + ky + lz) + i\phi_{hkl}]$$

The electron density distribution is best resolved by the fourier transform, if structure factors with high indices, resulting from high reflection angles, are included into the summation.

According to *Bragg's* equation,

$$\sin\Theta = \frac{\lambda}{2d} \qquad \frac{1}{d^2} = \frac{h^2}{a^2} + \frac{k^2}{b^2} + \frac{l^2}{c^2}$$

a, b, c are the unit cell edges

small distances *d* of crystal net planes, synonymous with high resolution, cause large diffraction angles, equal to high reflection indices (as represented here for the simple orthorhombic case).

The detection of reflections in a diffraction experiment is not time-resolved. Therefore, the direct measurement of reflection phase shifts is not possible with common crystallographic methods. Only the amplitudes of structure factors are obtained – they are the square roots of the measured reflection intensities I_{hkl}. The lack of direct experimental phases has been a great obstacle for structure determination in the past, and it still proves to be a challenge for macromolecular crystallography. It is known as the *Crystallographic Phase Problem*.

2.1.2 The experimental phasing of macromolecules with heavy-atom derivatives

There are several approaches to derive phases from reflection intensities. The so-called *Direct Methods* (*e.g.* Karle & Hauptmann 1956, not explained in detail) are purely statistical. Starting from a random set of phases and some basic assumptions, phase relationships are determined and exploited to derive complete phase sets. As the probabilities of these relationships are inverse-proportional to the square root of the number of atoms involved, and based on a sufficiently high resolution (with a limit of about 1.2 Å, Morris & Bricogne 2003), the method is restricted to small molecule structures – or small macromolecules for which highly resolved data has been collected.

If "heavy" atoms with an electron number significantly higher than that of the usual biological macromolecule atoms (carbon, oxygen, nitrogen, sulfur, phosphorus) are present in the structure, their contribution to the total scattering can be relatively large. Furthermore, in a map calculated directly from reflection intensities, their relative positions cause peaks from which the absolute positions can be derived. Such a map is called a *Patterson* map, and it is calculated from a special fourier transform, the *Patterson function*:

$$P(uvw) = \frac{1}{V}\sum_{hkl} \left|F_{hkl}\right|^2 \exp\left[-2\pi i(hu + kv + lw)\right]$$

The Patterson map is no ordinary electron density map. The peaks are not related to absolute atom positions, i.e. the origin of the patterson cell is not the crystal unit cell origin (although the cell dimensions are the same). As the Patterson function is calculated without phases, the coordinates are only relative ones, representing the interatomic distance vectors between the heavy atoms. Every pair of atoms forms a vector and thus a peak, also each atom with itself. These self-vectors have zero length and cause one common peak at the origin of the patterson cell. Apart from the origin peak (*N*-fold superposition), there are $N^2 - N$ peaks in a patterson map for *N* atoms. The Patterson peak heights are proportional to the product of the respective atomic numbers. For two atoms *A* and *B*, the vectors *A-B* and *B-A* have the same length, but the opposite direction (sign). Therefore, the Patterson function is centrosymmetric (even if the real structure is not), and its mathematic expression can be simplified because the sine terms are equal to the cosine terms:

$$P(uvw) = \frac{2}{V}\sum_{hkl} |F_{hkl}|^2 \cos[2\pi(hu + kv + lw)]$$

The Patterson function is valuable for heavy atom substructure determination, because the absolute heavy atom positions can be derived from the relative Patterson peak positions by solving suitable equation systems. Heavy atoms often tend to lie on special positions (0 or ½ on every cell edge), which simplifies the interpretation of the peaks. The great advantage of the Patterson function is its independence from phases. From the known heavy atom positions, phases $\phi_{H,calc}$ can be calculated and implemented in residual fourier synthesis, based on the difference between observed structure factor amplitudes, $|F_{obs}|$, and calculated ones for the heavy atoms, $|F_{H,calc}|$:

$$\Delta\rho(xyz) = \frac{1}{V}\sum_{hkl} \left(|F_{obs}| - |F_{H,calc}|\right)\exp\left[-2\pi i(hx + ky + lz) + i\phi_{H,calc}\right]$$

If the contribution of heavy atoms to the total scattering is dominating, the complete structure can in principle be solved by the heavy atom phases alone, because the remaining difference electron density peaks are precise enough to be interpreted. This is the case for most small molecule structures, where the (automated) patterson interpretation is directly followed by the structure refinement.

There is a drawback of this method for macromolecule structure determination. The scattering contribution of heavy atoms, although significant and useful in structure solution (see later), is not sufficient to apply the residual fourier synthesis and solve the structure directly. For example, a set of four mercury atoms (80 electrons each) in a 40 kDa protein (about 3000 atoms or 21,000 electrons) has only an electron contribution of about 1.5%. If many heavy atoms are incorporated into the derivative structure, like in case of the soaking method, their scattering contribution rises, but the patterson map becomes less interpretable, because the number of peaks increases and overlaps of peaks become more likely. If, for example, 20 heavy atoms are present in a structure, the number of extra-origin peaks is 380.

Still, every *experimental* macromolecular structure solution method is based on the determination of heavy atom substructures, from which macromolecule phases are derived in several ways. The basic principles of the different methods will be explained first in the following chapters; afterwards, some aspects of the single steps – substructure solution, macromolecule phase calculation and electron density improvement will be presented. Annotation: From now on, the expression “protein” is used as synonym for “macromolecule”, even if biological structures are not necessarily proteins.

2.1.2.1 Isomorphous Replacement

The term “Isomorphous Replacement” denotes the method of introducing heavy atoms into a protein structure without significantly changing the crystal geometry, i.e. the cell constants or even the crystal system. Diffraction data of a native protein crystal and (at least) one derivative protein crystal are collected. With the important precondition of isomorphism fulfilled, the structure factors of the heavy atom derivative F_{PH} are the (vector) sum of the native protein structure factors F_P and the separated heavy atom (sub-) structure factors F_H:

$$\vec{F}_{PH} = \vec{F}_P + \vec{F}_H$$

Protein structures are never centrosymmetric, but for spacegroups like $P2_12_12_1$, reflection projections perpendicular to the screw axes are centrosymmetric. The phase angles of the resulting centrosymmetric reflections are 0 or 180° (depending on the origin, 90° or 270° are also possible), *i.e.* their vectors lie on the real axis of the complex plane, and only the structure factor amplitudes have to be regarded:

$$|F_{PH}| = |F_P| \pm |F_H|$$

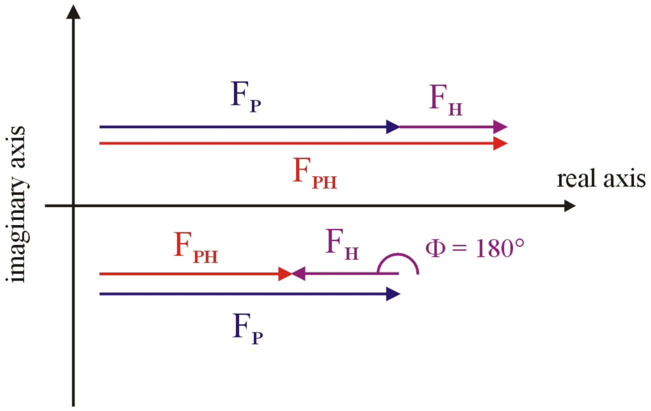


Fig. 2.2: Structure factor vector relationships for centrosymmetric reflections in the isomorphous replacement case.

Assuming that F_P and F_{PH} have the same sign, i.e. that F_H is smaller than F_P, the squared amplitudes of the heavy atom structure factors alone, $|F_H|^2$, can be derived from the amplitudes $|F_P|$ and $|F_{PH}|$, which are available as square roots of the measured intensities:

$$|F_H|^2 = (|F_{PH}| - |F_P|)^2$$

From the squared amplitudes, a patterson map of the heavy atom substucture can be calculated. Reflections with $F_P < F_H$ *and* opposite phase angles are rare enough not to distort the map significantly.

For non-centrosymmetric reflections, the isomorphous difference $\Delta|F|_{iso} = |F_{PH}| - |F_P|$ can be used as well, although it is not equal to $|F_H|$ in this case:

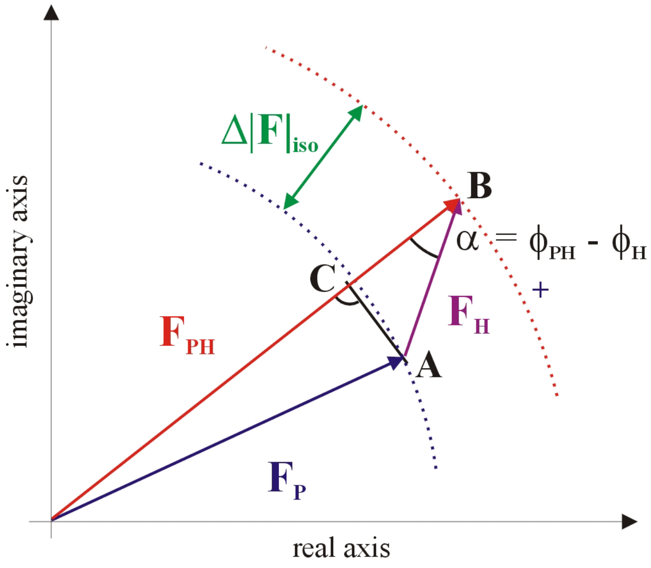


Fig.2.3: The structure factor vector triangle for the isomorphous replacement case. For the rectangular triangle *ABC* with hypothenusis $|F_H|$ it can be shown trigonometrically, that

$$\overline{BC} = |F_H|\cos(\alpha)$$

As generally $|F_H| << |F_{PH}|$ and $|F_P|$, the phase angle difference α is small and $\Delta|F|_{iso}$ is very similar to the straight *BC*, therefore:

$$\Delta|F|_{iso} \cong |F_H|\cos(\alpha)$$

Using the squared isomorphous differences in the Patterson function, the following expression emerges:

$$\Delta|F|_{iso} = |F_H|^2\cos^2(\alpha) = \frac{1}{2}|F_H|^2 + \frac{1}{2}|F_H|^2\cos(2\alpha)$$

Consequently, the Patterson function based on $\Delta|F|_{iso}$ for non-centrosymmetric reflections is on half the scale of the one based on centrosymmetric $|F_H|$ values, and it is more noisy because of the "useless" cosine term.

If the Patterson map has been successfully interpreted by a substructure solution program (e.g. SHELXD, details see later), the heavy atom phases are known from the positions, as well as the amplitudes of F_H, F_P and F_{PH} (the latter have been known before). Still, the phases of the native protein and the derivative are missing. They can in principle be deduced from the available information, applying a *Harker* construction (Harker 1956, fig. 2.4).

A circle of radius $|F_P|$ is drawn around an arbitrary origin. From the same origin, the vector $-F_H$ is drawn, and from the end point of this vector another circle, this time with radius $|F_{PH}|$. The two intersections between the two circles are finally connected to the circle centers. One set of obtained vectors corresponds to the correct structure factors F_P and F_{PH}. To break the ambiguity between the two possibilities for each structure factor, a *Single Isomorphous Replacement (SIR)* experiment is not sufficient. A second (different) derivative can be used to solve the probem. In this case, the method is called *Multiple Isomorphous Replacement (MIR)*.

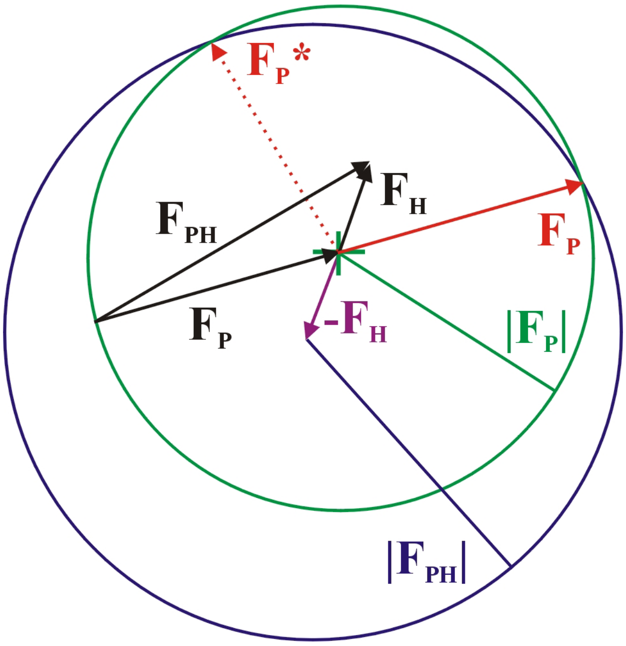


Fig.2.4: Harker construction for the SIR case. The black structure factor vector triangle is only displayed for orientation. F_P is the correct phase, F_P* a wrong solution. The construction of the two F_{PH} vectors from the intersections is not displayed.

2.1.2.2 Anomalous Scattering and the MAD experiment

The definition of the atomic scattering factors f_a, as presented before, is based on the assumption that atomic dispersion can be explained by a classical elastic scattering model. However, this assumption is only true for the theoretical free electron state. In reality, electrons are bound to atoms and located in electron shells. If the wavelength of interacting X-radiation is close to an absorption edge of a certain element (e.g. the K_α-absorption edge of copper), an inelastic contribution to the scattering process has to be taken into account. This is due to the fact that a fraction of X-Ray energy quantums is absorbed, causing electrons of the K shell to be ejected into the energy continuum. This first effect of inelastic scattering has the consequence that X-ray reflection amplitudes are altered. The temporarily “free” electrons may fall back into the K shell with a small time delay, emitting radiation of the original wavelength, but with a phase shift. This second effect causes differences in amplitude *and* phase for the diffracted X-rays.

The whole phenomenon is called *Anomalous Dispersion* (or anomalous scattering). It is usually neither relevant for light atoms like carbon, oxygen or nitrogen, nor for heavier elements, if the X-ray wavelength is far away from their absorbtion edges. But *if* an atom type is affected from anomalous dispersion, its atomic scattering factor evidently becomes anomalous:

$$f_{ano} = f_0 + f'(\lambda) + \; if''(\lambda)$$

As described by the equation, f_{ano} can be seperated into three components. The first one, f_0 corresponds to the normal elastic scattering contribution, which is wavelength-independent. The inelastic components f' and f'' are functions of the wavelength, f' being real and corresponding to partial X-ray absorption. It is also called the *dispersive signal* in an anomalous dispersion experiment. It only affects the scattering factor amplitude. The imaginary component f'', also called the *anomalous signal*, is influencing both amplitude and phase. The mathematical reason for this is the imaginary factor *i,* being equal to a 90° counter-clockwise phase angle shift in the complex vector plane (fig 2.5). The physical reason is, as mentioned before, the time-delayed re-emission of absorbed X-radiation.

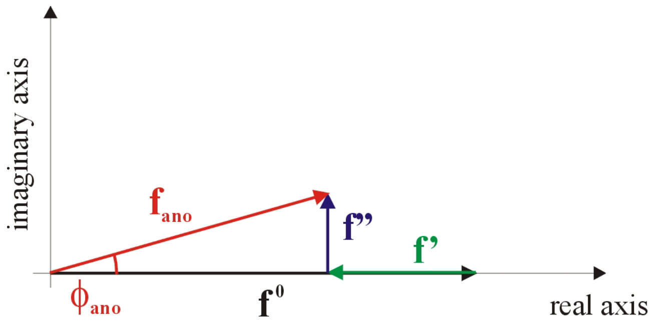


Fig.2.5: The connection between anomalous atomic scattering factor components, represented as vectors in the complex plane.

Anomalous dispersion has consequences for structure factors as well. *Friedel*'s law,

$$|F_{hkl}| = |F_{-h-k-l}| \qquad \phi_{hkl} = -\phi_{-h-k-l}$$

can be regarded as valid both for centrosymmetric and non-centrosymmetric structures, as long as anomalous dispersion is neglegible. If the effect becomes significant *and* the structure is acentric, *Friedel*'s law does not hold because of the anomalous amplitude and phase differences implied in the atomic scattering factors contributing to *F*. This fact can be exploited for protein phase determination with heavy atom derivatives – a heavy atom is also an anomalous scatterer, provided that a suitable wavelength is chosen.

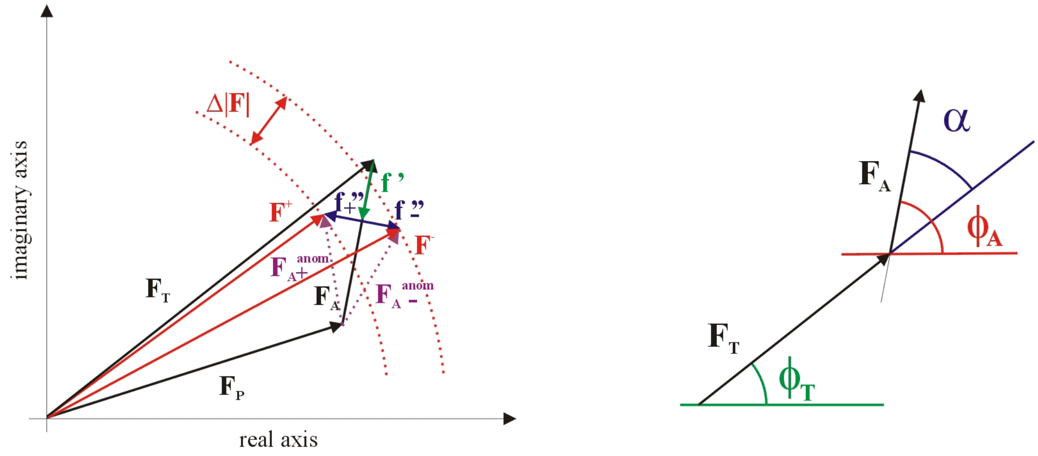


Fig. 2.6: Complex vector diagram of a derivative structure with significant anomalous contribution (one wavelength case). The construction explains the inequality of amplitude and phase for the resulting structure factors F^+ and F^-. Right: Illustration of the phase difference angle $\alpha = \phi_T - \phi_A$.

Looking at the usual structure factor vector triangle (fig. 2.6 left), F_P is not affected by anomalous scattering, because the native protein does not contain selenium or other heavy atoms – the anomalous signal of sulfur is weak, and neglegible at the selenium edge wavelength. F_A, the structure factor related to the anomalous scatterers, but ignoring the anomalous contribution, is connected to F_A^{anom}, the anomalous heavy atom structure factor, by f' and f''. This can be expressed qualitatively as $F_A^{anom} = F_A + f' + if''$. Similar to atomic scattering factor case (fig. 2.5), the real component f' does not change the F_A phase, but reduces its amplitude, because the f' value is negative (see scattering curve, fig. 2.7). The addition of the anomalous component f'' implies a 90° phase angle shift in either direction depending on the reflection orientation (f''_+ or f''_-). The vector construction finally leads to the anomalous derivative structure factors F^+ and F^-, for which the (inequal) amplitudes and their difference $\Delta|F|$ are available from the measured intensities.

In practice, selenium is the element that is still used most frequently for *multiple anomalous dispersion* (MAD) experiments. While the electron number of selenium is rather small compared to elements like iodine and mercury, and thus the scattering contribution is not very suitable for the *SIR* or *MIR* method, the anomalous dispersion at a wavelength around 0.98 Å is strong enough to be exploited for phasing. Furthermore, selenium can be incorporated into the protein by replacing the amino acid Methionine with Seleno-Methionine during protein expression, so that multiple derivatization can be obtained.

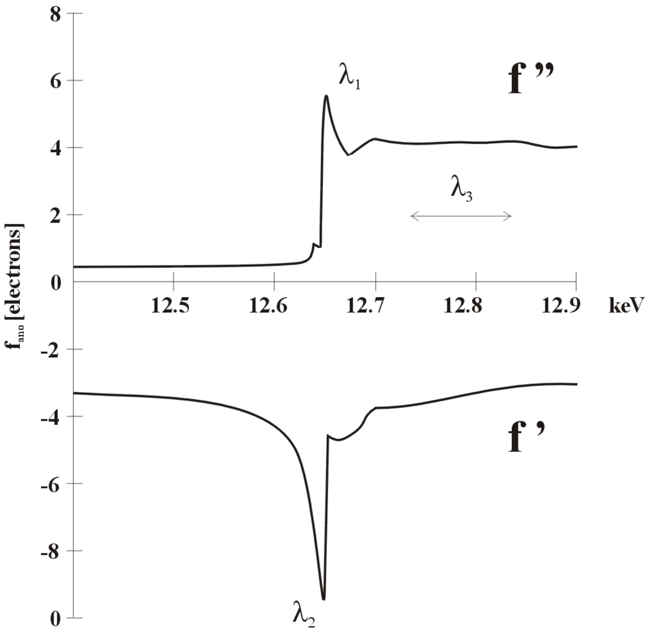


Fig. 2.7: The anomalous scattering curves (f ’ and f ”) for Selenium in the X-Ray energy region around the absorption edge. The absolute values are maximal at very similar wavelengths, the f ” maximum is at λ_1 = 0.980 Å (peak), the f ’ minimum is at λ_2 = 0.979 Å (inflection point). The data collection at least at these two wavelenths is known as the MAD (multiple anomalous dispersion) method. In a MAD experiment, a third data collection is often added in the so called high energy remote region (λ_3), where the f” component is high and does not change significantly with the wavelength. f” has to be measured by X-ray absorbtion spectroscopy, while f’ can be calculated from f”.

It has been shown (Karle 1980, Hendrickson et *al.* 1985) that the measured intensities of a MAD experiment can be mathematically described as follows:

$$\left|F^{\pm}\right|^2 = \left|F_T\right|^2 + \frac{f''^2 + f'^2}{f_0^2}\left|F_A\right|^2 + 2\frac{f'}{f_0}\left|F_T\right|\left|F_A\right|\cos\alpha \pm 2\frac{f''}{f_0}\left|F_T\right|\left|F_A\right|\sin\alpha, \qquad \alpha = \phi_T - \phi_A$$

In this equation, $|F_A|$, $|F_T|$, the derivative structure factor amplitude *without* anomalous contributions, and α, the difference phase angle between F_T and F_A, are unknown quantities. One wavelength is not sufficient to solve the equation, because there are only the two observables $|F^+|^2$ and $|F^-|^2$. However, using four observables per reflection from a two-wavelength MAD experiment or even six observables from a three-wavelength MAD, the equation system becomes over-determined and the missing quantities $|F_A|$, $|F_T|$ and α can be extracted from the equation.

The Argand diagram (Fig. 2.8) explains the same facts from a geometrical point of view. In particular, the availability of $|F_A|$ is important, because these amplitudes can be used for heavy atom substructure solution by Patterson (-aided) methods, like in the SIR / MIR case. It has to be emphasized that the two enantiomorph arrangements of the same given substructure are equally satisfying the patterson function and the direct methods solution. Already knowing angle α, the derivative phases, ϕ_T can be obtained from the calculated heavy atom phases by:

$$\phi_T = \phi_A + \alpha$$

With a fourier synthesis using $|F_T|$ and ϕ_T, an experimental electron density map for the protein derivative can be calculated, which should be interpretable after density modification (see later). If it is not, the inverted heavy atom substructure has to be taken to recalculate the phases.

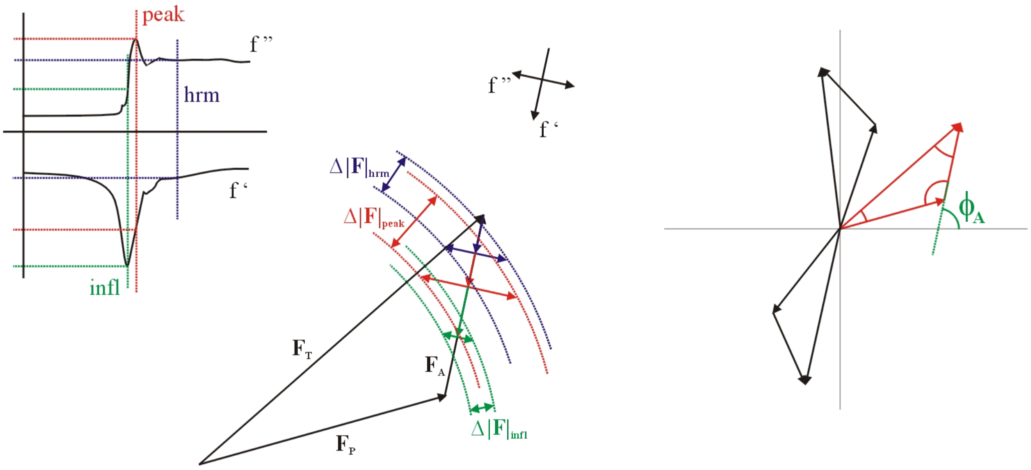


Fig. 2.8: The geometrical vector construction of a three-wavelength MAD experiment, using the f’ and f” contributions for peak, inflection point (infl) and high-energy remote (hrm) wavelength, as marked in the scattering curve (left). Together with the three Δ|F| values from the measured intensities, the trigonomical evaluation of this contruction yields all amplitudes and relative phase angle differences of the structure factor triangle. Yet, the absolute orientation of the triangle, i.e. the phase angles of either structure factor, are unknown (right). Determination of ϕ_A would solve the orientation problem, so that also the remaining phase angles could be derived.

2.1.2.3 SAD phasing and phase probabilities

Single anomalous dispersion (SAD) is the one-wavelength special case of MAD. It has already been emphasized that the restriction to two observables, $|F^+|^2$ and $|F^-|^2$, does not allow the solution of the MAD equation. Thus, neither $|F_A|$ nor the phase difference angle α can be determined exactly. Approximating these quantities, α can be set to 90° or 270° and the heavy atom substructure amplitudes $|F_A|$ replaced by the expression

$$|\Delta F| = \left(|F^+| - |F^-|\right)\,\sin\alpha \qquad \text{(valid if } |F_A| << |F_T| \text{ and } f', f'' << f_0\text{)}$$

i.e. $|\Delta F| = (|F^+| - |F^-|)$ for $\alpha = 90°$ $(|F^+| >> |F^-|)$ and $|\Delta F| = (|F^-| - |F^+|)$ for $\alpha = 270°$ $(|F^-| >> |F^+|)$

This replacement is not problematic for the substructure solution, if combined patterson / direct methods programs like SHELXD are applied (see the following chapter), because the large normalized structure factor amplitudes (E values) used in direct methods correspond to reflections with sin(α) close to ±1 anyway. With the heavy atom positions determined, calculated phases ϕ_A become available. Like in the MAD case, the protein derivative phase ϕ_T can be derived from the heavy atom phase by adding the difference angle α to it. However, as the two fixed α values are only approximations, the reliablity of ϕ_T depends on the question, how close the true phase differences are to 90° or 270°. Two cases can be distinguished:

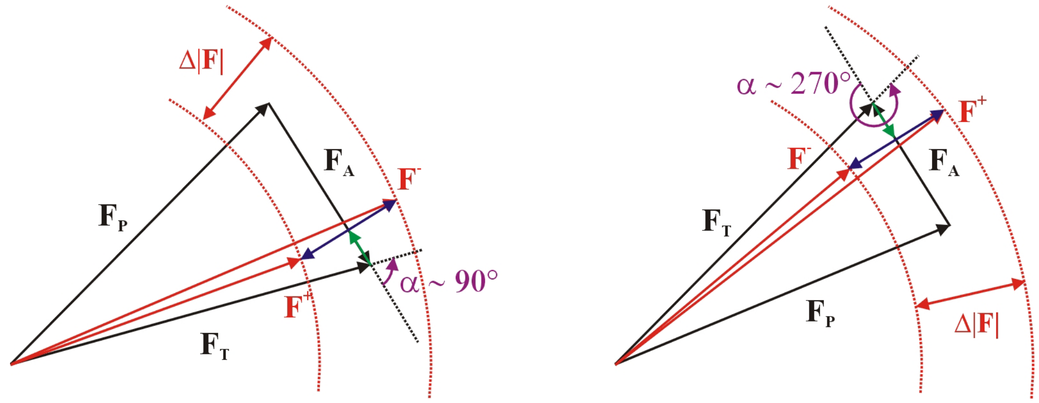


Fig. 2.9: SAD vector relationships in case of large Δ|F| values.

If Δ|F| is large, i.e if either $|F^+| >> |F^-|$ or $|F^-| >> |F^+|$, the α values are indeed close to 90° or 270°, respectively (Fig. 2.9). The two cases can be distinguished by the sign of Δ|F| (which determines also the assignment of α). In both cases, the addition of α to ϕ_A leads to a reliable phase ϕ_T.

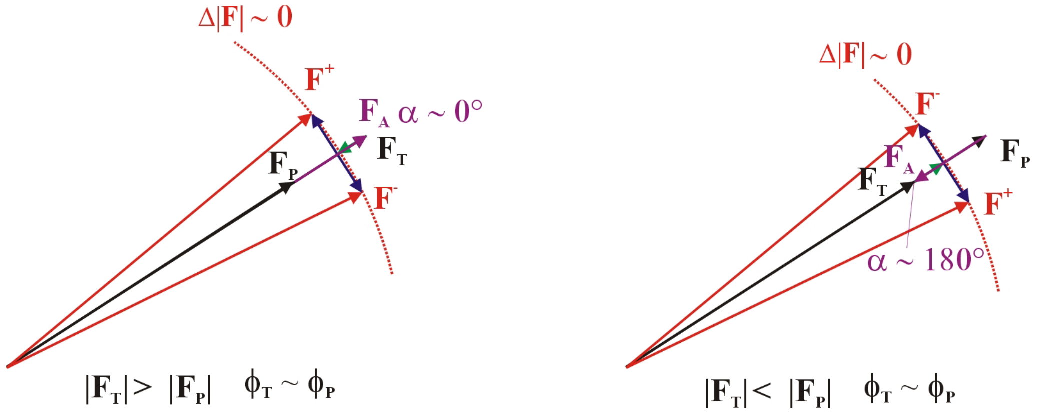


Fig. 2.10: SAD vector relationships in case of small Δ|F| values.

If Δ|F| is close to zero, i.e. if the measured anomalous structure factor amplitudes are almost equal, α is either close to 0° or close to 180°, depending on whether the amplitude of F_P or the one of F_T is larger (Fig. 2.10). In both cases, the phases of the two structure factors are very similar. It has to be remembered that the two cases can not be distinguished, because neither $|F_P|$ nor $|F_T|$ are known yet, and the phases of F^+ and F^- are unavailable. Apart from the two-fold ambiguity, the approximations of 90° or 270° for α have a maximum disagreement to reality (about ±90° phase error). Therefore, the addition of α to ϕ_A leads to a very unreliable phase ϕ_T.

Evidently, phase reliabilities of ϕ_T are very important for electron density calculation. The map accuracy will be increased, if reliable phases get a high weight and unreliable ones get a low weight. In every case of intrinsic two-fold ambiguity (e.g. SIR or SAD), there is an inevitable phase probability distribution, that has to be taken into account. But even for unambiguous phases, like in the MAD case, probability weights should be applied, because experimental phases are never error-free. For every phase of interest, the probability distribution function can be calculated. In ambiguous cases, the best phase to take is the so-called "centroid phase" $\phi(best)$, derived from the center of mass of the probability integral (Fig. 2.11). Also a quantity defining the normalized reliability of the centroid phase (and structure factor) can be derived. It is called the *figure of merit* (Blow & Crick 1959) and can be understood as the amplitude of the "reliability" vector $\underline{m}$ pointing to the probability centroid. The construction finally leads to a best-estimated structure factor $F_{hkl}(best)$ consisting of the measured amplitude $|F_{hkl}|$ for the given reflection and a phase term, based on $\phi(best)$ and weighted with the figure of merit m:

$$\vec{F}_{hkl}(best) = |F_{hkl}(best)| \exp[i\phi(best)] = |F_{hkl}| m \exp[i\phi(best)] \qquad \text{with } 0 < m = \frac{|F_{hkl}(best)|}{|F_{hkl}|} < 1$$

Since the centroid phase is the best estimate, but still not the true phase, the figure of merit downweights the centroid structure factor amplitude relative to the measured amplitude.

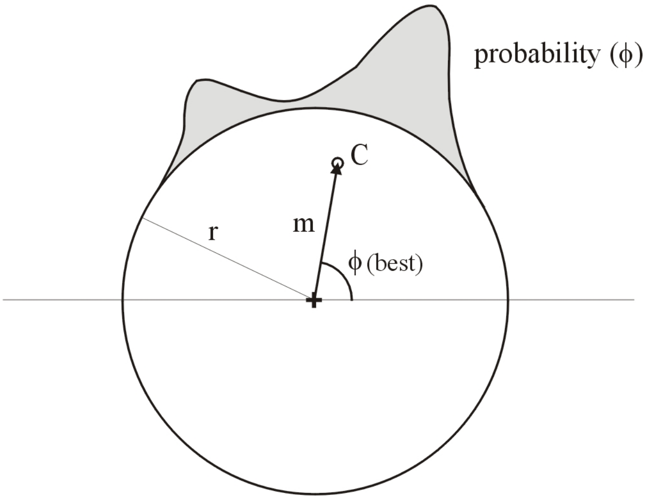


Fig. 2.11 Phase probability distribution in a case of two-fold ambiguity. The probability function P is placed on a circle of radius $r = 1$, which is related to a normalized overall probability for the structure factor of interest. The overall probability P of the corresponding (unknown) phase ϕ_{hkl} is the product of individual phase probabilities $P_i(\phi_i)$ along the probe section of the circle:

$$P(\phi_{hkl}) = \prod_{i=1}^{n} P_i(\phi_{hkl})_i$$

Transformed to the (structure factor) vector system, the product becomes a probability integral, for which the weighted average (the centroid C) can be determined. Vector $\underline{m}$ points to the centroid:

$$\vec{m} = \int_{\phi} (P_{hkl}(\phi)\exp[i\phi]) \; d\phi$$

The normalized centroid vector $\underline{m}$ can be understood as a kind of structure factor vector with centroid phase $\phi(best)$ and amplitude $|m|$, which is the figure of merit.

$$\vec{m} = |m| \exp[i\phi(best)] \qquad 0 < |m| = m < 1$$

In the example illustrated here, the two probability maxima correspond to possible solutions of a Harker construction (e.g. the SIR case of fig. 2.4), which are relatively close and with rather high local probabilities. Therefore, the centroid phase does not deviate much from either phase possibility (one of which is the true phase), and the figure of merit is quite high.

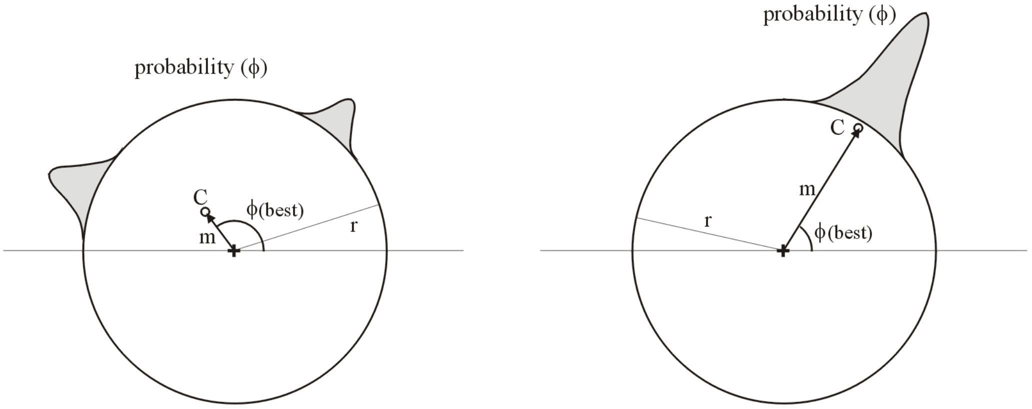


Fig. 2.12: Two contrary examples for overall phase probability. **Left:** Two-fold ambiguous stiuation, where the phase possibilities are far from each other and with a low local probability. The centroid phase deviates much from either solution and has a very low figure of merit. **Right:** An non-ambiguous phase with a high probablity (e.g. a MAD phase). The centroid vector is pointing to the probablity maximum without phase deviation. Still, the figure of merit is not quite one, because the experimental phase is not error-free (it still has a Gaussian probability distribution, although of narrow shape).

2.1.3 Practical aspects of experimental macromolecule phasing

2.1.3.1 Substructure solution with SHELXD

Using the previously determined heavy atom structure factor amplitudes $|F_A|$ (or their substitutes $|\Delta F|$ in the SAD case), as well as the (difference) phase angles α, the program SHELXD (Usón & Sheldrick 1999, Schneider & Sheldrick 2002) applies *patterson-aided direct methods* to determine the heavy atom positions, called *sites*. Refering to the *Shake-and-Bake* method (Miller et *al.* 1994), from which parts of the principle have been taken over (Sheldrick et *al.* 2001), the process can be called a “halfbaked” dual-space recycling algorithm. The method works as follows:

A start set of phases is generated from random atoms or from atoms at positions consistent with the $|F_A|^2$-based Patterson function. The phases are then refined and expanded in reciprocal space using *E*-value-based direct methods, in particular the *tangent formula* (Karle & Hauptmann 1956). From the refined phases, electron density is calculated by a *Fast-Fourier-Transform* (FFT) algorithm. In the real space density map, the *N* strongest peaks are picked (*N* being the approximately expected number of sites for which similar scattering power and a minimum distance is assumed), while weak peaks related to noise are ignored. Structure factors are then re-calculated from the positions of the picked peaks, and the phase refinement step is repeated. Usually, for a set of *N* random starting atoms, *2N* cycles of alternating dual-space steps are performed. During the last few cycles, site occupancies are refined using the peak heights of the final electron density map.

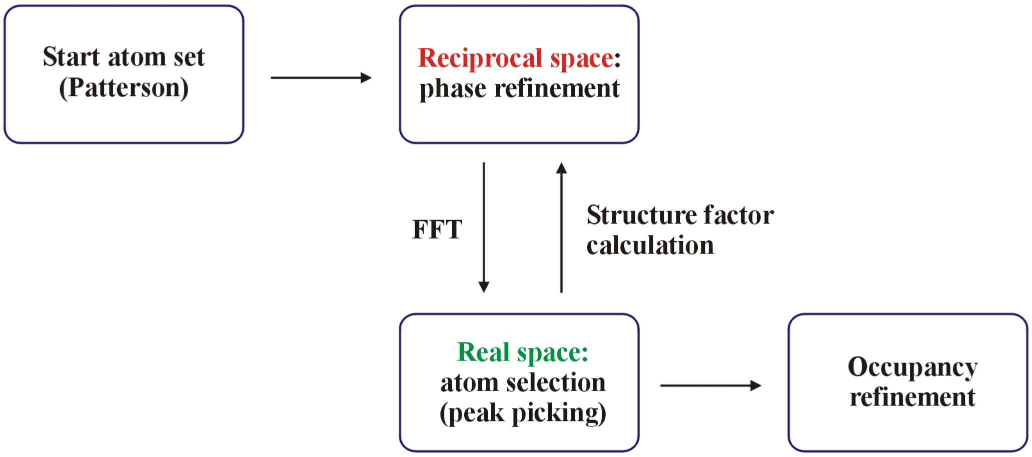


Fig. 2.13: Flow chart of SHELXD operations for heavy atom substructure determination

The heavy atom substructure solution trials are validated using a reciprocal space correlation coefficient (Fujinaga & Read 1987) for the agreement between E_o and E_c, the observed and calculated normalized structure factors.

$$CC = 100\frac{\left[\sum_{hkl}(wE_oE_c)\sum w - \sum_{hkl}(wE_o)\sum_{hkl}(wE_c)\right]}{\sqrt{\left[\sum_{hkl}(wE_o^2)\sum w - \left(\sum_{hkl}(wE_o)\right)^2\right]\left[\sum_{hkl}(wE_c^2)\sum w - \left(\sum_{hkl}(wE_c)\right)^2\right]}}$$

Additionally, crossword tables containing peaks of the Patterson superposition function are produced for the best solutions (according to the CC), allowing manual identification and assignment of reasonable heavy atom sites.

2.1.3.2 Substructure refinement and protein phase calculation with SHARP

The program SHARP (de La Fortelle & Bricogne 1997) derives protein derivative phases ϕ_T from a previously determined heavy atom substructure (F_A, ϕ_A) and the measured structure factor amplitudes (e.g. several values for $|F^+|$ and $|F^-|$ in the MAD case). The heavy atom parameters – coordinates, occupancies and B-values – as well as the anomalous dispersion components *f'* and *f''* are (optionally) refined to improve the heavy atom phases and thus also the accuracy of the derived protein phases. After each major refinement step, residual electron density maps are calculated in order to find (possible) additional heavy atom sites. The residual map results from the difference between observed structure factor amplitudes $|F_{obs}|$ and calculated ones, related to the already known structure information – preliminary protein structure factors F_P and (incomplete) heavy atom model structure factors F_H*.

$$\Delta\rho(xyz) = \frac{1}{V}\sum_{hkl}\left(\left|F_{obs}\right| - \left|F_P + F_{H*}\right|\right)\exp\left[-2\pi i(hx + ky + lz) + i\phi_{calc}\right]$$

The heavy atom model refinement of SHARP uses the *Bayesian maximum likelihood* formalism (Bricogne 1991). The principle of maximum likelihood (also called maximum entropy) is based on *Bayes' Theorem* for joint probabilities (Pannu & Read 1996). For example, two quantities A and B are considered, which have probabilities *p* based on the truth of each other. Thus, be *p(A; B)* the probability of *A,* given that *B* is known to be true, and be *p(B; A)* the probability of *B,* given that *A* is true. Then, according to Bayes' Theorem, the joint probability *p(A,B)* for the two quantities follows a multiplicative law:

$$p(A,B) = p(A)\,p(B;A) = p(B)\,p(A;B)$$

The joint probability is the product of the independent probability of one quantity with the conditional probability of the other quantity.

Transfered to crystallography, the two quantities are the observed data (A), and the model parameters (B). The model is the quantity to be optimized, so in terms of model refinement, one is interested in *p(model; data)* – the probability of the model, given the data. From Bayes' Theorem, one obtains

$$p(model;\ data) = \frac{p(model)\,p(data;\ model)}{p(data)},$$

where *p(data),* the probability of the data alone, can be regarded as constant, since the data are fixed. Unlike to this, *p(data; model),* the (theoretical) probability to observe these data, given that a certain model would be true, is not predefined. Taking into account that the model parameters *x* are independent, the probability of the model (given the data) can be replaced by the likelihood *L*, being the product of the individual parameter probabilities: $L = \Pi\, p(x_i)$.

From the last two considerations, it follows that

$$L\ (model;\ data) = p(model)\ p(data;\ model)$$

In the crystallographic model likelihood definition, the data-independent model probability *p(model)* is not arbitrary. Some models are prefered to others *a priori* due to a physically more reasonable geometry. Therefore, *p(model)* is given by geometry restraints, expressing an a-priori probability.

For reasons of numerical convenience, the log-likelihood *LL* is prefered to *L*, replacing the product of probabilities by a summation: $LL = \ln(L) = \sum_{i=1}^{n} \ln[p(x_i)]$

Model parameter refinement with Bayesian methods is trying to maximize the log-likelihood target function, therefore the name “maximum likelihood”. The advantage of the method compared to its special case of least-squares refinement, is that a gaussian error distribution for the parameters is not required and also systematic errors are allowed. In practice, this facilitates the refinement of more imperfect (incomplete) models with less resolved data.

2.1.3.3 Density modification and solvent flattening

In cases where the “raw” protein phases obtained after the initial phasing process have low reliabilities, the errors of the resulting electron density are too large for map interpretation. Thus, the protein structure can not yet be regarded as solved. There are methods to apply general prior knowledge about the structure in order to improve the phases. These methods are comprized as *density modification*.

If the asymmetric unit of the crystal cell contains several copies of the macromolecule, related to each other by more or less exact symmetry (*non-crystallographic symmetry, NCS*) the corresponding regions of preliminary electron density can be averaged to minimize density errors and to improve the phases after re-inversion of the map. This method is called *NCS averaging* (Vellieux & Read 1997). It requires the determination of the *NCS* rotation matrix and the so-called protein density mask, describing the shape of electron density that belongs to a single molecule copy. *NCS* averaging is for example implemented in the program *DM* (Cowtan 1994).

Another density modification approach exploits the fact that the electron density distribution is different for the protein region and the “bulk” solvent region of the cell (usually 30 – 70% of the cell volume). For the protein, the average density as well as the density variance is higher than for the solvent.

Technically, the two regions are distinguished as follows: A grid is placed into the asymmetric unit of the cell. The standard deviation of electron density $\sigma(\rho)$ is calculated for spheres (with radius $r \sim 3$ Å) around every grid point (*pixel*). The mean electron density in the sphere, $<\rho>$ and the individual electron density values ρ_i for *N* pixels lying in the sphere are contributing to σ:

$$\sigma(\rho) = \sqrt{\frac{\sum_{i=1}^{N} (\rho_i - \bar{\rho})^2}{N}}$$

The grid points are sorted by the height of the σ values. The list is devided into two fractions corresponding to the percentage of protein and solvent in cell (which has been approximately determined before). The lower fraction is regarded as solvent region. For all solvent pixels, the density can be set to a low constant value, so that noise peaks are removed and the density distribution becomes flat. The method is therefore called *solvent flattening*. The technique of inverting a solvent density weighting factor γ is called *solvent flipping*. It is applied in some density modification programs such as SOLOMON (Abrahams & Leslie 1996) or SHELXE (Sheldrick 2002). SHELXE, a combined protein phasing and density modification program, "flips" solvent pixels ($\rho_S{}^* = -\gamma\rho$ where $\gamma \sim 1$) and modifies positive electron density in the protein region by replacing ρ with

$$\rho_P^* = \sqrt{\frac{\rho^4}{\rho^2 + g^2\sigma^2(\rho)}} \qquad \textit{(g is usually 1.0)}$$

The method of *histogram matching* (Lunin 1988) approximates a given distribution of poor electron density to a theoretical distribution (histogram). This technique is based on the fact that the theoretical protein density distributions are found to be characteristic for a given data resolution and solvent content, but independent from the nature of the specific protein.

2.2 Macromolecular structure refinement

The idea of structure refinement, being the final step of a crystal structure determination, is the optimization of an initial molecule model obtained after the phase problem solution. This model, no matter whether it results from *Molecular Replacement* or has been built (manually or automatically) into an experimental electron density map, is likely to be incomplete and has many parameter errors, for example inaccurate atom coordinates. The refinement of parameters is done in such a way that the structure factor amplitudes calculated from the model $|F_{calc}|$ approximate the observed structure factor amplitudes $|F_{obs}|$. The agreement between both (for all F_{hkl}) is given by the R-factor:

$$R = \frac{\sum_{hkl} \left\| F_{obs} \right| - \left| F_{calc} \right\|}{\sum_{hkl} \left| F_{obs} \right|}$$

With the model optimization, the R-factor is minimized. Evidently, also the calculated phases improve with increasing model accuracy.

To perform the parameter refinement, a target function Q has to be defined. In case of the *least-squares* method, used by the refinement program SHELXL (Sheldrick & Schneider 1997), Q is the sum of squared errors and has to be minimized. For SHELXL, the error is the deviation of squared structure factor amplitudes, $\Delta_2 = |\Delta|F|^2| = ||F_{obs}|^2 - |F_{calc}|^2|$, therefore:

$$Q = \sum_{hkl} w_{hkl} \left(|F_{obs}|^2 - |F_{calc}|^2 \right)^2$$

w is a weighting factor, by which less accurate structure factors are downweighted.

The target function is at a minimum, if the structure factor sum Σ_{hkl} of partial function derivatives for $|F_{calc}|^2$ against the individual model parameters p_i becomes zero. In the mimimum-condition equation,

$$\sum_{hkl} w_{hkl} \left(|F_{obs}|^2 - |F_{calc}|^2 \right) \frac{\partial F_{calc}^2}{\partial p_i} = 0\,,$$

the substitution of F_{calc} by all its separated partial derivatives,

$$F_{calc}^2 = F_{c(0)}^2 + \frac{\partial F_c^2}{\partial p_1} \Delta p_1 + \frac{\partial F_c^2}{\partial p_2} \Delta p_2 + \ldots \frac{\partial F_c^2}{\partial p_n} \Delta p_n\,,$$

leads to a system of normal equations, represented in vector notation,

$$\begin{bmatrix} a_{11} & a_{21} & a_{31} & \ldots \\ a_{12} & a_{22} & a_{32} & \ldots \\ a_{13} & a_{23} & a_{33} & \ldots \\ \ldots & \ldots & \ldots & \ldots \end{bmatrix} \times \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \ldots \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \ldots \end{bmatrix},$$

where $\underline{\varepsilon}$ is a vector of parameter shifts, $\underline{b}$ is the *gradient vector* containing the sum of structure factor deviations $\Sigma w_{hkl}\Delta|F|^2(\partial|F_{calc}|^2/\partial p_i)$, and $[A]$ is a symmetric square matrix containing all combinations $[i, j]$ of partial derivatives, the elements being $\Sigma w_{hkl}\ (\partial|F_{calc}|^2/\partial p_i)(\partial|F_{calc}|^2/\partial p_j)$.

The refinement is an iterative process, where the parameter variation should lead to minimization of the gradient vector and to convergence of the paramater shifts after a certain number of cycles. For macromolecular structures, the full matrix refinement implies a considerable amount of calculations. To reduce the calculation time, the off-diagonal matrix elements can be set to zero. This simplification is practicable, because the diagonal matrix elements of $[A]$ are larger than the rest. The most frequently used algorithm for the iterative least squares minimization with the diagonal matrix $[A]_0$ is the *conjugate gradient technique* (Tronrud 1992). It increases the rate of convergence introducing initial estimates of parameter shifts ε_0, used in a residual matrix $[\boldsymbol{R_0}] = \underline{b} - [A]\ x\ \underline{\varepsilon}_0$. The target function Q is minimized along a search direction vector.

A more general target function, the *log-likelihood LL*, can be used in an alternative refinement approach. The underlying principles and the advantages compared to the least squares special case have been presented in chapter 2.1.3.2. The program REFMAC (Murshudov et *al* 1997) uses maximum likelihood refinement.

The model parameters to be refined are at least the atomic coordinates x, y and z. Depending on the number of reflection data available, more parameters related to atomic displacements can be introduced. If the data resolution is low, a global isotropic B-value for the whole protein or several B-values for groups of residues are used. Also the coordinates may be fixed for very low resolution cases, refining the position of one or more *rigid bodies* only. At higher resolution (below a limit of about 2 Å), individual isotropic B-values can be used. The number of parameters for N atoms becomes $4N$. Very high resolution data (about 1.2 Å or higher) allows the refinement of six anisotropic displacement parameters (*ADP*s) per atom, leading to a total number of $9N$ parameters.

To avoid under-determination of the refinement problem, a sufficiently high data-to parameter ratio is required. In general, the use of prior knowledge about molecular geometry, applied in form of several distance *restraints*, can be regarded as an increase of the refinement data number. Also *ADP*s are usually restrained. The types of restraints used by SHELXL are shortly explained in the experimental methods section.

Certain definite parameters, such as the coordinates of atoms on special crystallographic positions, need not to be refined. Therefore, these *constraints* reduce the number of model parameters.

The final structure optimization process is done by the alternating steps of manual model corrections – the readjustment or replacement of atoms as well as the building of new model parts – and numerical parameter refinement cycles, as explained. The unavailability of a sufficient data number and the (neccessary) use of restraints instead of experimental data often implies the danger of over-interpretation of the model. This problem is favoured by the fact that model building relies on difference electron density maps resulting from *model-calculated* phases ϕ_{calc},

$$\Delta\rho(xyz) = \frac{1}{V}\sum_{hkl}\left(|F_{obs}| - |F_{calc}|\right)\exp\left[-2\pi i(hx + ky + lz) + i\phi_{calc}\right]$$

Therefore, the agreement of the model to the difference map as well as to observed structure factor amplitudes may tend to be "self-fulfilling". The problem is known as *model bias*. To avoid it, the crystallographic *Cross-Validation method* (Brünger 1992) has been developed. The idea is to exclude a (usually random) set of e.g. 5% of reflections, the *free set* of structure factor amplitudes, from the refinement. After the refinement against the remaining 95% of reflections, the *working set*, the model may be biased towards $\Sigma_{hkl}|F|^2_{work}$, but not towards $\Sigma_{hkl}|F|^2_{free}$, so the corresponding R_{free} index is lower than R_{work}. As long as the difference between both structure quality indicators does not become too large, the model building and refinement can be regarded as safe from over-interpretation. To avoid bias in electron density maps, a *Sigma-A* weighted map, $2mF_{obs} - DF_{calc}$, should be used in addition to the normal difference density map (Read 1986 & 1990).

3 Materials and Methods

3.1 Studies on the crystal structure of human Aldose Reductase

3.1.1 Expression, purification and crystallization of Aldose Reductase

Human Aldose Reductase (hAR2) was expressed in *E. coli* with a $(his)_6$-tag, purified by metal-affinity chromatography and co-crystallized with the oxidized form of the coenzyme ($NADP^+$) and the inhibitor IDD594 at pH 5.0 and 277 K, after thrombin cleavage of the tag (Lamour et *al.* 1999). The crystals belong to the space group $P2_1$ with unit cell dimensions a = 49.43 Å, b = 66.79 Å, c = 47.40 Å and $\beta = 92.4°$, one complex per asymmetric unit and a solvent content of 34.6%.

3.1.2 Data collection and reduction (Howard et *al.*, submitted to proteins)

Three-wavelength MAD data were collected at a APS synchrotron beamline from one single crystal of the seleno-met derivative hAR2-IDD594 complex, diffracting to 0.90 Å. Data reduction was carried out using the HKL2000 package (Otwinowski & Minor 1997) with the programs DENZO for integration and SCALEPACK for scaling. The data subsets from different wavelengths were treated independently.

3.1.3 Data analysis and exploitation

The scaled data were obtained in three separate scalepack files for the wavelength subsets with already merged intensities of symmetry equivalent reflections. To create a pseudo-native data set for refinement as well as an anomalous data set for phasing, the data files were treated with the program XPREP (Bruker AXS, Madison, U.S.A).

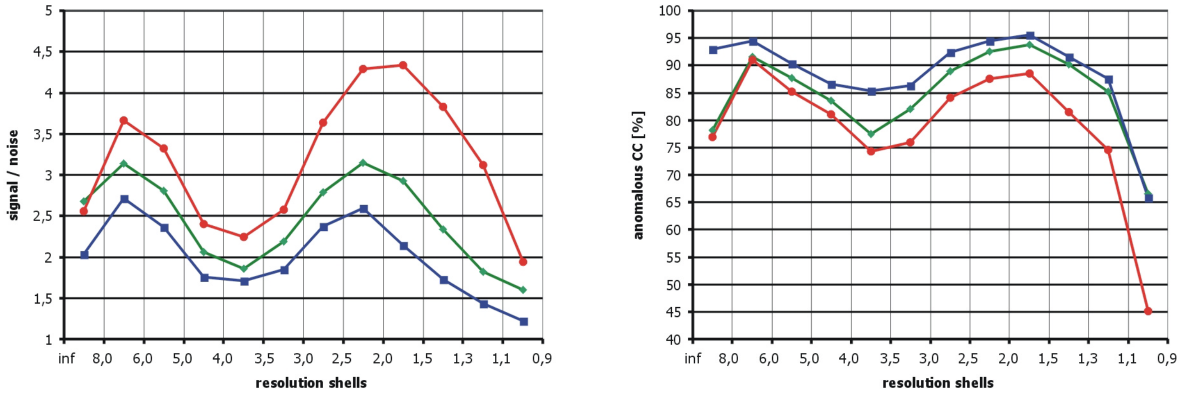
First, the primitive monoclinic cell geometry and the resulting space group $P2_1$ were confirmed by checking systematic absences of intensities due to the presence of translating symmetry operators. Given the cell constants and the necessary chirality of the structure, $P2_1$ was the only possible spacegroup. Analyzing the merged Friedel pair intensities for all three wavelength subsets, the high-energy remote data were found to have the highest quality and completeness (Table 3.1). They were therefore chosen as a pseudo-native data set for refinement. After flagging 5% of the reflections for cross-validation, a file containing the respective averaged structure factor intensities (SHELX *HKLF-4* format) was written.

To provide F_A data for SHELXD, signed anomalous differences (based on F^+ and F^-) were calculated from the diffraction intensities using XPREP. The derived F_A amplitudes and phase difference angles (see theory) were written into a single (SHELX *HKLF-3* format) file. The diagnostic program output was taken to analyze the precision and accuracy of the anomalous data as a function of the resolution (Fig 3.1).

data subset (wavelength)		high-en. remote (0.9465 Å)	peak (0.9793 Å)	inflection point (0.9795 Å)
symmetry-merged reflections		434,721	407,824	369,895
Friedel-merged-reflections		220,816	207,339	188,756
completeness [%]	all	97.3	91.2	82.9
	1.0-0.9 Å	90.8	68.2	57.2
intensity / sigma	all	18.2	17.4	14.8
	1.0 – 0.9 Å	9.7	7.2	5.4
R_{merge}[1] [%]	all	4.4	6.5	4.3
	1.0 – 0.9 Å	10.3	15.5	15.0
R_{sigma} [%]	all	3.4	3.9	3.9
	1.0 – 0.9 Å	9.5	13.9	18.4

Table 3.1: Statistics of the three wavelength data subsets. [1]Note that the reflection statistics are related to already symmery-averaged intensities, therefore R(merge) refers to the agreement of Friedel pairs only.

The peak data exhibit the highest signal-to-noise ratio for the whole resolution range. A prominent feature at all wavelengths is a local minimum of signal-to-noise at about 3.5 – 4.0 Å resolution. It still has to be investigated if this phenomenon can be explained by the noise of a so-called water ring. The correlation between the lack of data precision and the resulting phase quality will be discussed later. The graphs describing the correlation between signed anomalous differences for different combinations of wavelengths show a corresponding decrease of data accuracy in the so-called water ring region, and a steep fall-off beyond 1.1 Å resolution. Nevertheless, the correlation is well above the empirical limit of 30% (Schneider & Sheldrick 2002) for all cases and up to full 0.9 Å resolution.

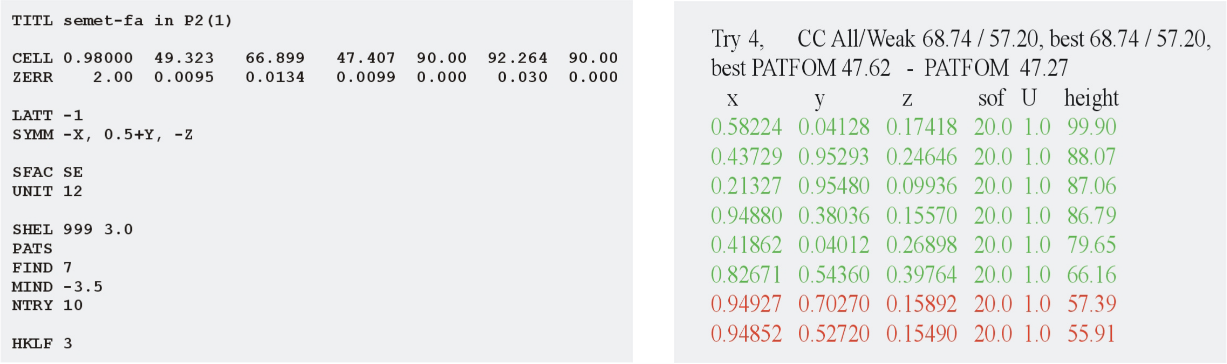


(a) **(b)**

Fig. 3.1: Quality of the anomalous hAR data. (a) Signal-to noise ratios for signed anomalous differences at the three wavelengths – peak (red), high energy remote (green) and inflection point (blue). (b) Correlation between each pair of wavelength subsets – peak vs. inflection (blue), peak vs. high energy remote (green), inflection vs. high energy remote (red).

3.1.4 The localization of selenium sites

The program SHELXD (Schneider & Sheldrick 2002) was used to solve the *hAR2* heavy atom substructure by determining the positions of the anomalously scattering selenium atoms. A Patterson-aided “halfbaked” dual-space algorithm was applied (see theory).



```
TITL semet-fa in P2(1)

CELL 0.98000  49.323   66.899   47.407  90.00  92.264  90.00
ZERR    2.00  0.0095   0.0134   0.0099  0.000   0.030  0.000

LATT -1
SYMM -X, 0.5+Y, -Z

SFAC SE
UNIT 12

SHEL 999 3.0
PATS
FIND 7
MIND -3.5
NTRY 10

HKLF 3
```

(a)

```
Try 4,    CC All/Weak 68.74 / 57.20, best 68.74 / 57.20,
best PATFOM 47.62  -  PATFOM  47.27
   x         y         z       sof   U    height
0.58224  0.04128  0.17418  20.0  1.0  99.90
0.43729  0.95293  0.24646  20.0  1.0  88.07
0.21327  0.95480  0.09936  20.0  1.0  87.06
0.94880  0.38036  0.15570  20.0  1.0  86.79
0.41862  0.04012  0.26898  20.0  1.0  79.65
0.82671  0.54360  0.39764  20.0  1.0  66.16
0.94927  0.70270  0.15892  20.0  1.0  57.39
0.94852  0.52720  0.15490  20.0  1.0  55.91
```

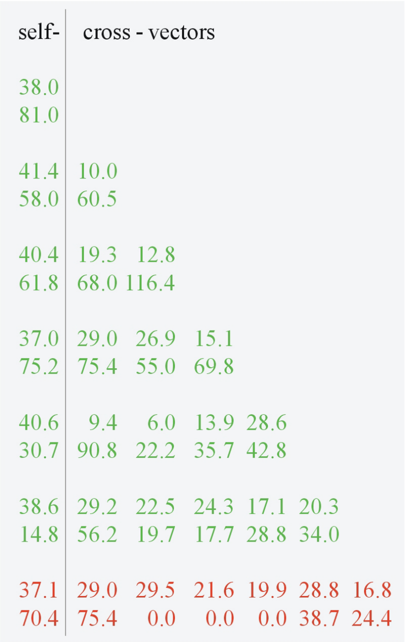
(b)

Fig.3.2: **(a)** Instructions for the SHELXD job against hAR F_A data truncated to 3.0 Å resolution**. (b)** Coordinates and Patterson peak heights of the anomalous scatterers of the best solution (right). SHELXD-2001, an early version of the program used during these studies, did not derive the selenium atom occupancies from the peak heights.

One molecule of the hAR2 selenium derivative contains 6 Se-methionines and one bromine atom belonging to the ligand. Running SHELXD in default mode against F_A data truncated to 3.0 Å (looking for 7 heavy atom sites), six reasonable sites were obtained. While the peak list for the best solution try # 4 (Fig. 3.2 b) did not show a clear step in peak height, the crossword table clearly supported six positions. The seventh possible heavy atom site produced three cross-vectors to previous sites, for which no patterson peak heights were observed (Fig. 3.3).

The respective SHELXD job was carried out with ten solution tries only. To further evaluate the effect of data quality on substructure accuracy, a test job was repeated using SHELXD-2003, the most recent program version, requesting 100 solutions tries. Like for the former job, where the CC(E_{obs}, E_{calc}) values had varied within the very narrow range of 67.9 – 68.7%, here the solution CC values were very close as well, between 74.7 and 75.5%. Unlike the earlier version, SHELXD-2003 determines heavy atom occupancies, which are refined by the conjugate-gradient least-squares method.

The analysis of solutions (Fig. 3.4) exhibits a unimodal distribution for pairs of correlation coefficient and PatFOM values, which is unusual as the distribution is normally bimodal. Regardless the version of SHELXD, the very high F_A data quality provides a 100% success rate of solutions.



self-	cross - vectors					
38.0						
81.0						
41.4	10.0					
58.0	60.5					
40.4	19.3	12.8				
61.8	68.0	116.4				
37.0	29.0	26.9	15.1			
75.2	75.4	55.0	69.8			
40.6	9.4	6.0	13.9	28.6		
30.7	90.8	22.2	35.7	42.8		
38.6	29.2	22.5	24.3	17.1	20.3	
14.8	56.2	19.7	17.7	28.8	34.0	
37.1	29.0	29.5	21.6	19.9	28.8	16.8
70.4	75.4	0.0	0.0	0.0	38.7	24.4

Fig. 3.3 Crossword table for the best solution (CC = 68.7%) obtained from the SHELXD-2001 job run on hAR2 F_A data

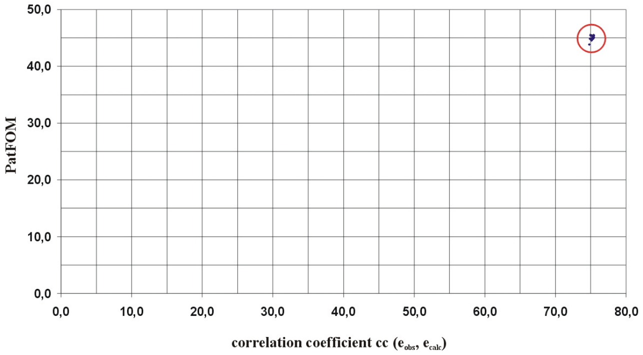


Fig. 3.4 Scatterplot of PatFOM values versus correlation coefficients of normalized structure factors, CC (E_{obs},E_{calc}), for 100 solutions obtained from the SHELXD-2003 job on hAR2 F_A data. The distribution is clearly unimodal and very narrow.

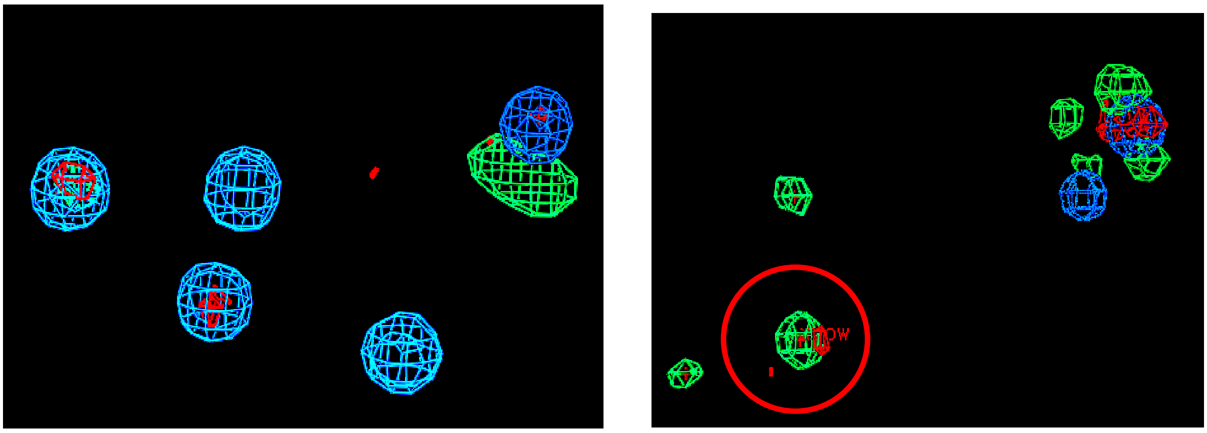
3.1.5 Heavy atom model refinement and protein phase calculation

The set of six anomalous scatterers obtained from SHELXD was refined using SHARP v.1.3.8 beta for Linux (de La Fortelle & Bricogne 1997) with standard settings, allowing anisotropic B-value and occupancy refinement for all heavy atoms in the later stages. The number of positionally fixed waters, as known from a previously refined hAR2 model, was included into the number of protein light atoms (light-atom F-fraction in SHARP).

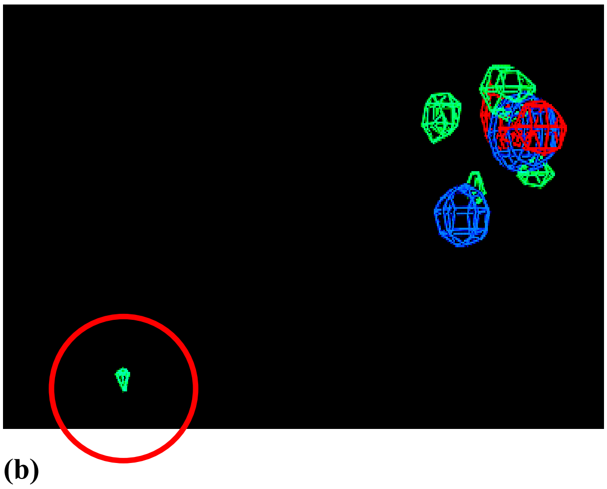
step	sites	B-values	resolution [Å]	remark
1	6 Se	isotropic	3.0	all sites starting with B = 30, occ. = 1.0
2	7 Se	isotropic	2.0	sites 6, 7 starting with occ. = 0.7 : 0.3
3	7 Se	isotropic	1.5	--
4	7 Se	anisotropic	1.2	sites 6,7 remain isotropic
5	7 Se	anisotropic	0.9	--

Table. 3.2: SHARP heavy atom refinement conditions in the phasing process for hAR2. To save computing time and to facilitate a stable refinement, the data resolution limit and the number of refined parameters were increased successively.

After the major refinement steps, it was tried to locate possible multiple conformations of the already modelled anomalous selenium scatterers and to find the missing bromine site with help of residual maps. Also, the visibility of anisotropy was analyzed to judge the necessity of anisotropic refinement. After the first refinement step at 3.0 Å, a residual peak of 13.8 σ close to site six was observed (Fig.3.5 a) and, a new partly occupied site was modelled at its position.



(a) **(c)**



(b)

Fig. 3.5: SHARP electron density maps (blue) and residual electron density maps (green and red) at different resolutions. **(a)** maps from anomalous peak data at 3.0 Å (the residual map contoured at 4 σ), showing a secondary selenium position. **(b)** and **(c)**: residual maps at 1.5 Å contoured at 7 σ, from anomalous peak and remote data, respectively. Only **(c)** clearly shows the bromine site (circle). Anisotropy effects are visible in both maps. Please note that the perspective of image **(a)** differs from **(b)** and **(c)**.

The bromine site present in the hAR2 structure was only observed in the residual map associated with the anomalous remote-wavelength data (Fig. 3.5 c). This phenomenon is explained by the different scattering curves of bromine and selenium (Fig. 3.6). At the peak of the anomalous selenium f'' curve, the corresponding f'' for bromine is still in the low energy pre-peak region, whereas at about 14 keV, the high-energy remote side of the f'' for Se, also the anomalous bromine signal is close to its peak. The Bromine site was not included into the SHARP heavy atom model.

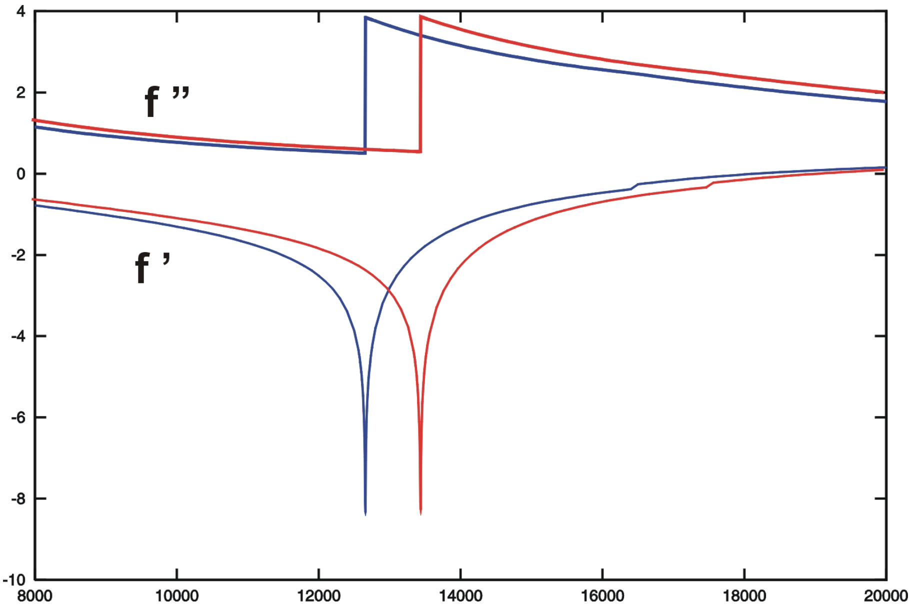


Fig. 3.6: Anomalous f' and f'' scattering curves as functions of wavelength (energy) for selenium (blue) and bromine (red). The X-ray enery is given in eV. The graphs were created by Ethan Meritt's web server (http://www.bmsc.washington.edu).

The final model for the selenium substructure contained seven sites, the last two corresponding to two alternative positions for the same Se atom (Table 3.3). Judging from its high B-values, site five seems to be disordered as well, but a clear secondary position could not be assigned from a residual peak. The disorder of the corresponding methionine residue was modelled later during the protein refinement. The disordered pair of positions six and seven reveals still high B-values and an occupancy sum of only 60%. This fact reflects the fexibility of the methionine side chain in the protein model, where the refinement of threefold disorder was attempted with limited success (see results).

Site	x	y	z	occ	B(iso)	B11	B22	B33
1	0.5803	0.0421	0.1743	0.9961	4.87	4.94	4.92	4.74
2	0.4381	0.9569	0.2470	0.8454	5.70	5.76	5.82	5.53
3	0.2127	0.9569	0.0981	0.9045	4.80	4.99	4.85	4.56
4	0.9486	0.3798	0.1569	0.9310	4.64	4.87	4.54	4.51
5	0.4214	0.0360	0.2686	0.8973	26.16	29.68	24.84	23.96
6	0.8262	0.5451	0.4000	0.4218	20.03	-	-	-
7	0.7814	0.5087	0.3817	0.1932	14.46	-	-	-

Table 3.3: The final parameters of the selenium substructure after the last SHARP refinement job. $B_{11,22,33}$ are the major anisotropic displacement parameters (diagonal matrix elements). In case of anisotropic B-values, B(iso) is the equivalent isotropic value: $B = 1/3\ \mathrm{trace}(\underline{B})$.

3.1.6 Density modification by solvent flattening

Phase improvement by solvent flattening is usually applied after the initial phase calculation from the substructure sites only. This phasing step enhances the quality of the electron density map, so that the tracing of the first protein model is simplified. In cases when high-quality MAD data is available, solvent flattening is in principle not necessary, because the phases are unambiguous and sufficiently reliable. The solvent-flattening procedure implies some model assumptions (see theory), so that the resulting map can not strictly be regarded as purely experimental. Thus, it is not suited in the context of these studies.

As a test for the phasing work on hAR2, the density modification program SOLOMON (Abrahams & Leslie 1996) was used to flatten the solvent part of the SHARP map. The program was started with default setting, a solvent content of 43% and 20 solvent flattening cycles. The differences between the phases obtained from SOLOMON and the raw SHARP phases were measured with SHELXPRO (see results) The solvent-flattened map was not further studied. The unmodified SHARP phases and the corresponding map were used for model verification after completing the hAR2 structure refinement.

3.1.7 Refinement of the hAR protein structure with SHELXL

With an experimental electron density of sufficient resolution and phase quality, most of the protein model can be automatically built. Depending on the resolution, the modeling of only the backbone or even the side chains of the protein is possible. Several programs exist for this task. However, in the scope of this work, the model building step was completely skipped in favour of using an existing model previously refined with SHELXL (Sheldrick & Schneider 1997) against the 0.66 Å native data (Howard et *al.*, submitted to proteins).

3.1.7.1 Refinement strategy overview

The model provided by the Podjarny group was used to initiate the SHELXL refinement. To reduce model bias, all water molecules and multiple conformations were removed and the atom positions were randomly modified on the order of 0.05 Å per atom coordinate. The new refinement was carried out using the Friedel-merged high-energy remote intensities with 5% of the data set aside for cross-validation. The refinement was performed using a standard protocol close the the one described by Sheldrick & Schneider (1997). The sequence of steps was as follows:

step(s)	remark	data	parameters	R_{work}	R_{free}
1	first SHELXL refinement of initial model	49055	10629	21.02	24.12
2-4	345 water oxygens atoms and 24 secondary conformations modelled	49055	12091	16.23	19.69
5-7	inclusion of data to 0.9 Å and anisotropic displacement parameters	220816	27201	12.69	14.52
8	190 water oxygen atoms modelled	220816	29166	10.67	12.24
9	non-polar hydrogen atoms (HFIX)	220816	29313	9.08	10.70
10	C-terminus (313-315) removed	220816	29465	8.98	10.60
11-23	more double conformations, free variables introduced	220816	30912	8.48	10.15
24-36	mainly half-occupied water oxygen atoms, conformational adjustments, some new conformations	220816	32358	7.93	9.54

Table 3.4: Key refinement steps (or step sequences) with resulting R-values. The column data lists the total number of Friedel-merged unique data, 95% of which were used for refinement and 5% for cross-validation.

In the first steps (1 to 4, see Table 3.4) only data to 1.5 Å were used and the atomic B values were treated isotropically. In this phase, most of the fully occupied solvent water oxygen atoms were set and the clearest double side chain conformations were modelled. In step five, the resolution was increased to 1.15 Å, rising the number of data from 49,055 to 108,687. After that, the number of model parameters was increased by including anisotropic displacement parameters into the refinement. This lead to a drop of 2.8% in R_{work} and 2.1% in R_{free}. In step seven, all 220,816 data up to 0.9 Å were included into the refinement. Another 190 fully occupied water oxygen atoms were added to the model in step eight. Finally, the addition of hydrogen atoms in step nine lead to a drop of 1.6% in R_{work} and 1.5% in R_{free}.

The following more detailed refinement steps improved the model further, but did not lower the R values drastically. Between steps 11 to 23, mostly side chain disorder, which had not become obvious before, was modelled. During this refinement phase, free occupancy variables were assigned to the occupancies of all multiple atom positions. After step 23, networks of disordered atoms were identified and modelled using common free variables. In the last steps of modeling and refinement, remaining peaks were interpreted as half- or otherwise partly occupied water oxygen atoms. Large deviations of stereochemical properties from target values (restraint violations) were systematically checked and used to correct and adjust the placement of disordered atoms. Aspects of the various modeling steps are explained in more detail during the following sub-chapters.

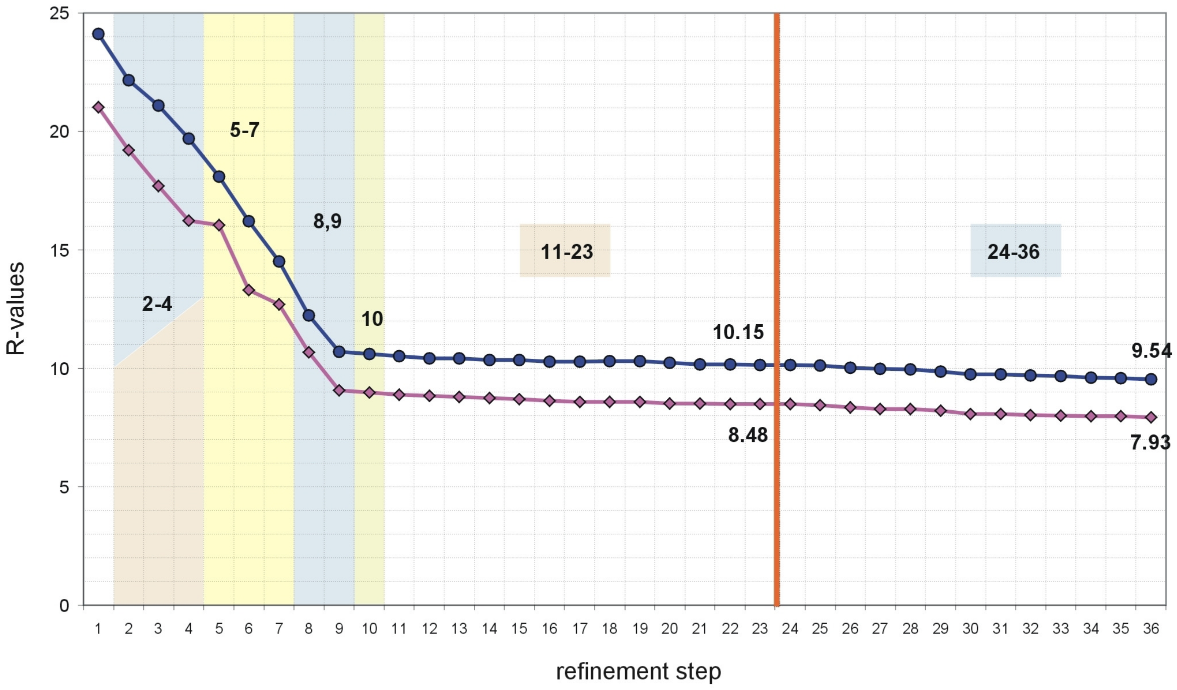


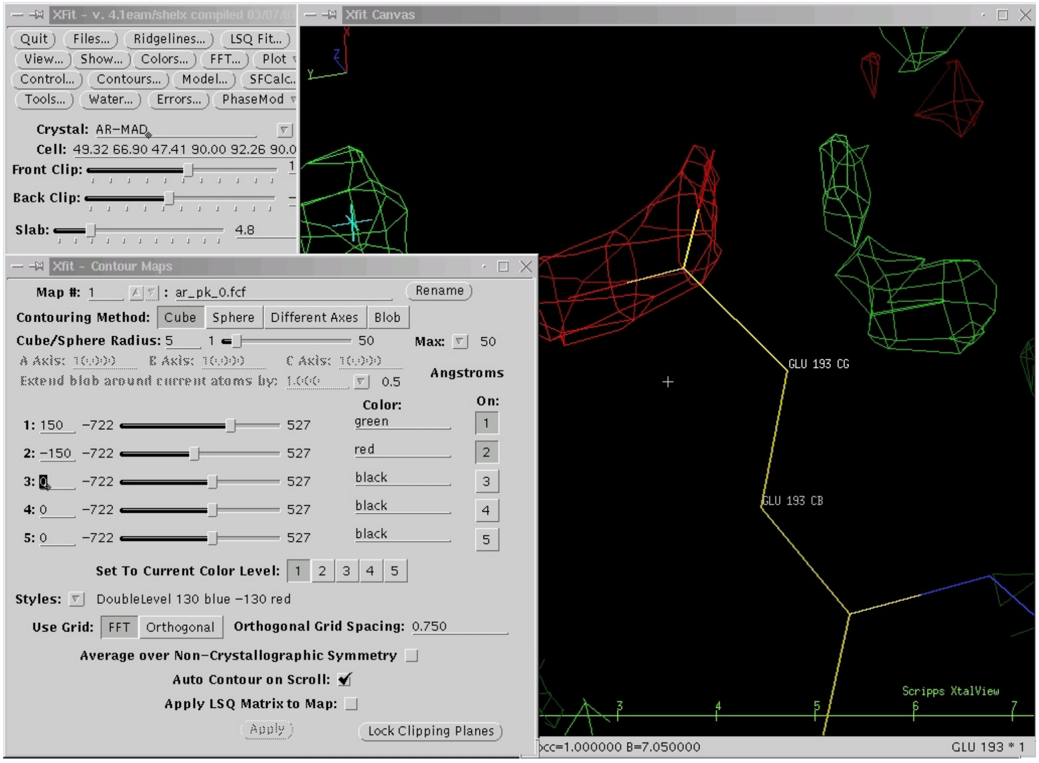
Fig. 3.7 The development of the R-values during the refinement process. Blue curve for R_{free}, pink curve for R_{work}. Water modeling steps are coloured in light-blue, disorder-modeling steps in light pink, data/parameter increasing steps in light yellow and the introduction of hydrogen atoms in light green.

3.1.7.2 Model refinement with SHELXL

During the hAR model improvement process, the refinement jobs were computed with SHELXL using the conjugent gradient method with ten refinement cycles per job (*CGLS 10*). Model parameters were refined against reflection intensities (F^2). To ensure a stable refinement in the first cycle of each calculation, the multiplicative parameter shift factor for CGLS was set to 0.4 (*DAMP*, the default value is 0.7). 200 difference electron density peaks with a minimum distance of 2.3 Å to other atoms or peaks were listed at the end of each job (*PLAN 200 2.3*, only *100* in the later stages). Anomalous differences were ignored in scattering factor calculation for F_c (*MERG 4*), in agreement to the use of the Friedel-merged data. Geometrical restraints were applied to 1,2- and 1,3-distances (*DFIX, DANG*), to the planarity of amide groups, guanidine groups in arginine residues and aromatic ring systems (FLAT) and to the chirality of asymmetrically bonded atoms like Cα (*CHIV*). Anisotropy restraints were used to assimilate the overall displacement ellipsoid directions (*SIMU*) and the bonded atom pair displacements parallel to their bond directions (*DELU*). For *DFIX, CHIV, FLAT* and *DELU*, the recommended restraint esd values were kept (defaults were explicitly set with DEFS 0.02 0.1 0.01). The default of 0.04 for *SIMU* was raised to 0.1, allowing a less rigid refinement of anisotropic displacement, taking into account the high data-to-parameter ratio and the low isotropic B-values for most protein atoms. To limit anisotropy for water oxygen atoms, their anisotropic components were restrained to approximate isotropic behaviour (*ISOR*). Anti-bumping restraints were applied to avoid distances smaller than expected non-bonded values (*BUMP*). All restraints were applied in all refinement steps. Atomic occupancies were refined with individual free variables for all independent groups of disordered atoms. A matrix inversion for the determination of standard deviations of the parameters has not yet been done.

3.1.7.3 Map display and modeling with XTALVIEW / XFIT

For a modeling session with XFIT (McRee 1999), the most recent PDB file from SHELXL was used as model source. The corresponding FCF type phases file was read twice to generate a σ_A-weighted map (coefficients $2mF_o$-$1DF_c$) and a F_o-F_c difference electron density map. For normal modeling tasks, FFT-calculated maps with auto-contouring were used. In cases where alternative conformations were presumed but the density was unclear, the positive difference level was reduced to 2.5 or even 2 σ (125/ 100) and, if necessary, the σ_A level was reduced to 0.7 σ (35). In the first four modeling sessions, all residues were systematically inspected for side chain disorder or misplaced side chains.



Crystal: AR-MAD
Cell: 49.32 66.90 47.41 90.00 92.26 90.0
Front Clip:
Back Clip:
Slab: 4.8

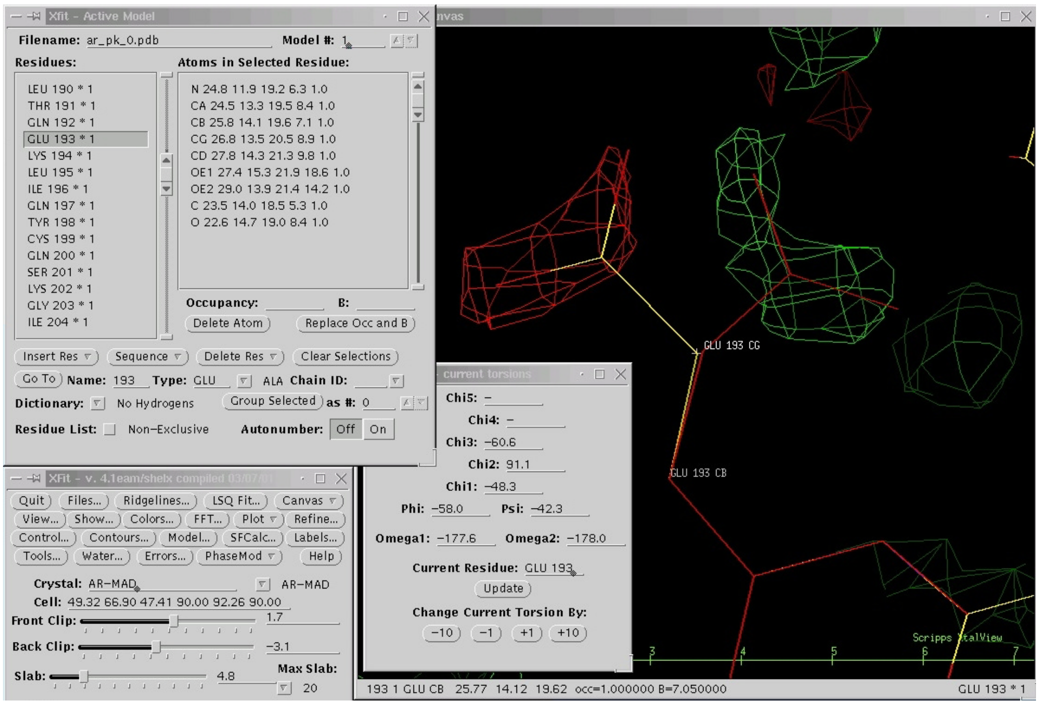
Map #: 1 : ar_pk_0.fcf
Contouring Method: Cube Sphere Different Axes Blob
Cube/Sphere Radius: 5 1 50 Max: 50
A Axis: 10.000 B Axis: 10.000 C Axis: 10.000
Extend blob around current atoms by: 1.000 0.5 Angstroms

				Color:	On:
1:	150	−722	527	green	1
2:	−150	−722	527	red	2
3:	0	−722	527	black	3
4:	0	−722	527	black	4
5:	0	−722	527	black	5

Set To Current Color Level: 1 2 3 4 5
Styles: DoubleLevel 130 blue −130 red
Use Grid: FFT Orthogonal Orthogonal Grid Spacing: 0.750
Average over Non-Crystallographic Symmetry ☐
Auto Contour on Scroll: ☑
Apply LSQ Matrix to Map: ☐

occ=1.000000 B=7.050000 GLU 193 * 1

Fig. 3.8: XFIT setting for a typical Fo-Fc difference electron density map. The program uses 50 units for one sigma. Map contours were set to single level of 1 sigma (50) with blue color for the σ_A map and to double level of 3 σ (150) with green color and –3 σ (-150) with red color for the Fo-Fc map.



Filename: ar_pk_0.pdb Model #: 1

Residues:

LEU 190 * 1
THR 191 * 1
GLN 192 * 1
GLU 193 * 1
LYS 194 * 1
LEU 195 * 1
ILE 196 * 1
GLN 197 * 1
TYR 198 * 1
CYS 199 * 1
GLN 200 * 1
SER 201 * 1
LYS 202 * 1
GLY 203 * 1
ILE 204 * 1

Atoms in Selected Residue:

N 24.8 11.9 19.2 6.3 1.0
CA 24.5 13.3 19.5 8.4 1.0
CB 25.8 14.1 19.6 7.1 1.0
CG 26.8 13.5 20.5 8.9 1.0
CD 27.8 14.3 21.3 9.8 1.0
OE1 27.4 15.3 21.9 18.6 1.0
OE2 29.0 13.9 21.4 14.2 1.0
C 23.5 14.0 18.5 5.3 1.0
O 22.6 14.7 19.0 8.4 1.0

Occupancy: B:
Name: 193 Type: GLU ALA Chain ID:
Dictionary: No Hydrogens as #: 0
Residue List: ☐ Non-Exclusive Autonumber: Off On

Chi5: –
Chi4: –
Chi3: −60.6
Chi2: 91.1
Chi1: −48.3
Phi: −58.0 Psi: −42.3
Omega1: −177.6 Omega2: −178.0
Current Residue: GLU 193
Change Current Torsion By:

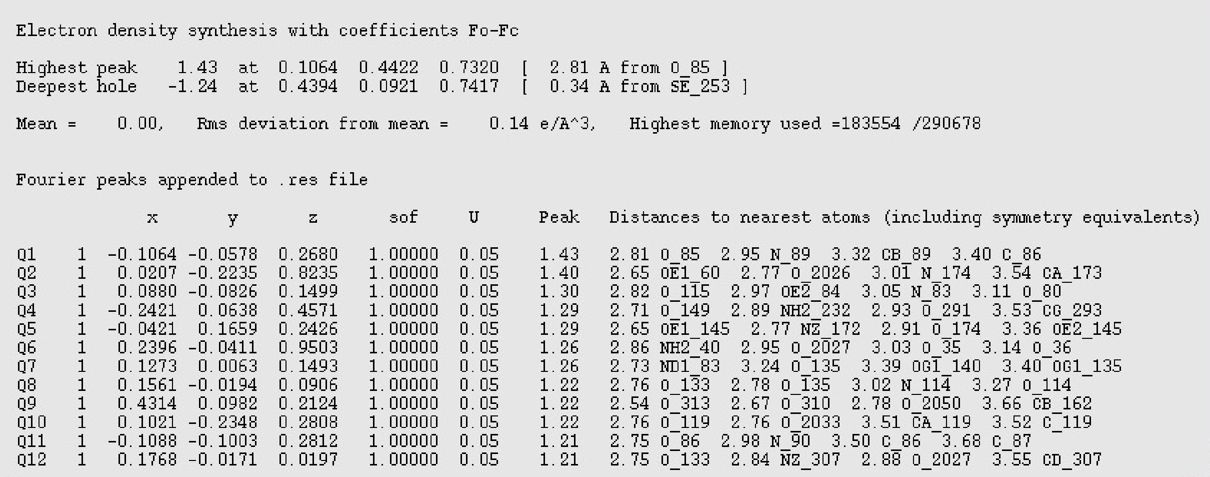
Crystal: AR-MAD AR-MAD
Cell: 49.32 66.90 47.41 90.00 92.26 90.00
Front Clip: 1.7
Back Clip: −3.1
Slab: 4.8 Max Slab: 20

193 1 GLU CB 25.77 14.12 19.62 occ=1.000000 B=7.050000 GLU 193 * 1

Fig. 3.9: Sidechain modeling procedure using XFIT. Sidechain conformations were modelled after selecting the respective residue and rotating along the side chain bonds, thus altering the torsion angles. New conformations were just fixed (*apply fit*) or set as additional secondary conformations (*split sidechain*, then *apply fit*) in case of disorder.

3.1.7.4 The refinement of solvent water molecules

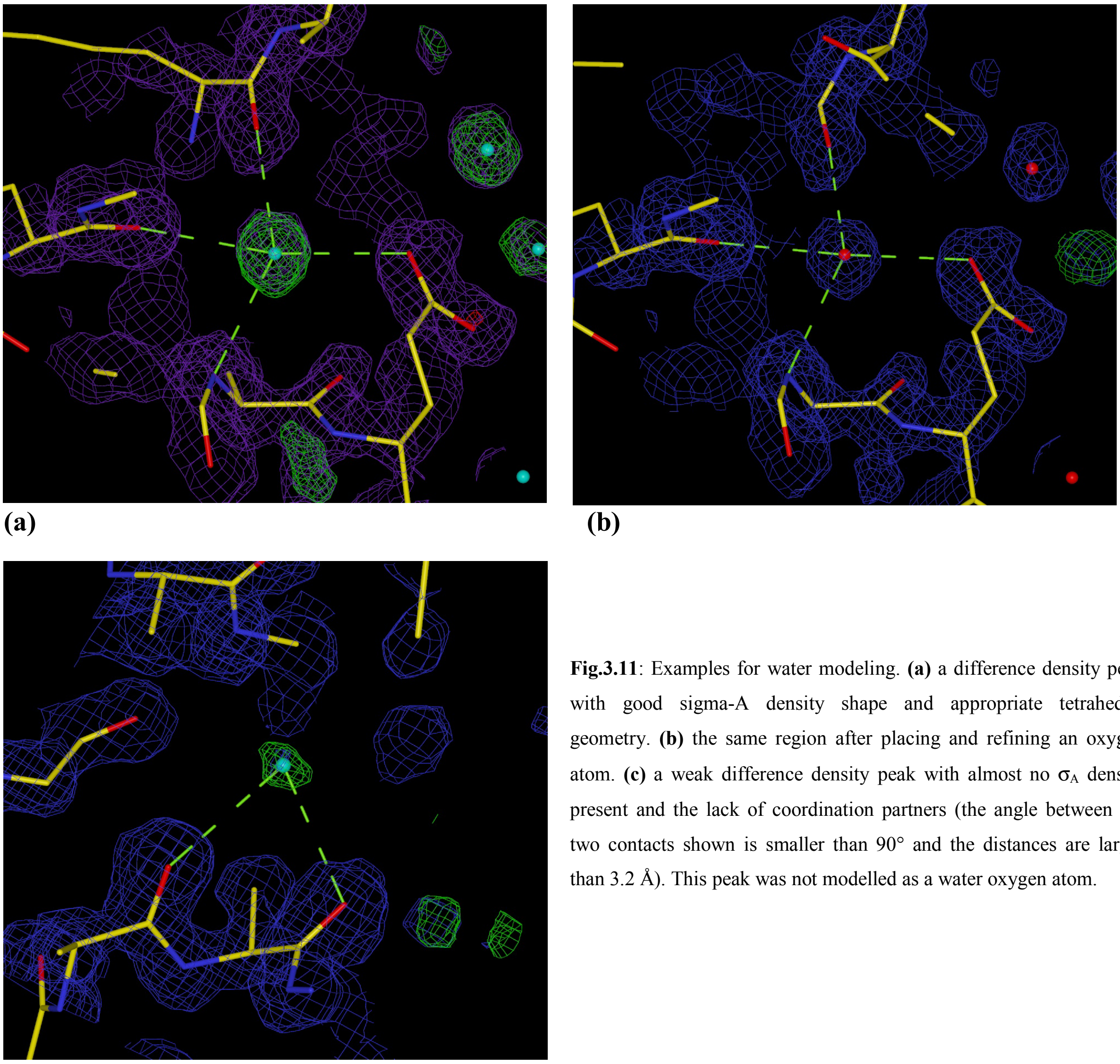
After the last refinement cycle of each job, SHELXL calculates an F_o-F_c electron density map and lists a certain number of highest local maxima (200 by default) in the protocol file, given with their coordinates, peak heights (in e/Å^3) and neighbouring atoms.



```
Electron density synthesis with coefficients Fo-Fc

Highest peak    1.43  at  0.1064  0.4422  0.7320  [  2.81 A from O_85 ]
Deepest hole   -1.24  at  0.4394  0.0921  0.7417  [  0.34 A from SE_253 ]

Mean =    0.00,   Rms deviation from mean =    0.14 e/A^3,   Highest memory used =183554 /290678


Fourier peaks appended to .res file

             x       y       z       sof      U      Peak   Distances to nearest atoms (including symmetry equivalents)

Q1    1  -0.1064 -0.0578  0.2680  1.00000  0.05    1.43   2.81 O_85  2.95 N_89  3.32 CB_89  3.40 C_86
Q2    1   0.0207 -0.2235  0.8235  1.00000  0.05    1.40   2.65 OE1_60  2.77 O_2026  3.01 N_174  3.54 CA_173
Q3    1   0.0880 -0.0826  0.1499  1.00000  0.05    1.30   2.82 O_115  2.97 OE2_84  3.05 N_83  3.11 O_80
Q4    1  -0.2421  0.0638  0.4571  1.00000  0.05    1.29   2.71 O_149  2.89 NH2_232  2.93 O_291  3.53 CG_293
Q5    1  -0.0421  0.1659  0.2426  1.00000  0.05    1.29   2.65 OE1_145  2.77 NZ_172  2.91 O_174  3.36 OE2_145
Q6    1   0.2396 -0.0411  0.9503  1.00000  0.05    1.26   2.86 NH2_40  2.95 O_2027  3.03 O_35  3.14 O_36
Q7    1   0.1273  0.0063  0.1493  1.00000  0.05    1.26   2.73 ND1_83  3.24 O_135  3.39 OG1_140  3.40 OG1_135
Q8    1   0.1561 -0.0194  0.0906  1.00000  0.05    1.22   2.76 O_133  2.78 O_135  3.02 N_114  3.27 O_114
Q9    1   0.4314  0.0982  0.2124  1.00000  0.05    1.22   2.54 O_313  2.67 O_310  2.78 O_2050  3.66 CB_162
Q10   1   0.1021 -0.2348  0.2808  1.00000  0.05    1.22   2.76 O_119  2.76 O_2033  3.51 CA_119  3.52 C_119
Q11   1  -0.1088 -0.1003  0.2812  1.00000  0.05    1.21   2.75 O_86  2.98 N_90  3.50 C_86  3.68 C_87
Q12   1   0.1768 -0.0171  0.0197  1.00000  0.05    1.21   2.75 O_133  2.84 NZ_307  2.88 O_2027  3.55 CD_307
```

Fig.3.10 Part of a SHELXL lst file containing the difference electron density maxima. The line before the peak list shows the density rms level corresponding to 1 σ. For example with the rms level being 0.07 e/Å^3, peak Q1 of the list with a height of 0.63 e/A^3 has a relative level of 9 σ.

Assuming a relatively complete model lacking only some secondary conformations, the highest remaining peaks usually correspond to water oxygen atoms. The modeling of these atoms was done with XFIT, displaying both the σ_A and the F_o-F_c maps and sequentially checking the model for peaks. The possible water oxygen atoms were judged by three criteria: a sufficient relative peak height in the difference density map, i.e. more than five sigma, the presence of a spherically shaped σ_A density contour (at 1 σ) at the same position and the existence of ideally four hydrogen bond partner atoms (donors or acceptors), leading to a tetrahedral coordination geometry for the respective atom.



(a) **(b)**

(c)

Fig.3.11: Examples for water modeling. **(a)** a difference density peak with good sigma-A density shape and appropriate tetrahedral geometry. **(b)** the same region after placing and refining an oxygen atom. **(c)** a weak difference density peak with almost no σ_A density present and the lack of coordination partners (the angle between the two contacts shown is smaller than 90° and the distances are larger than 3.2 Å). This peak was not modelled as a water oxygen atom.

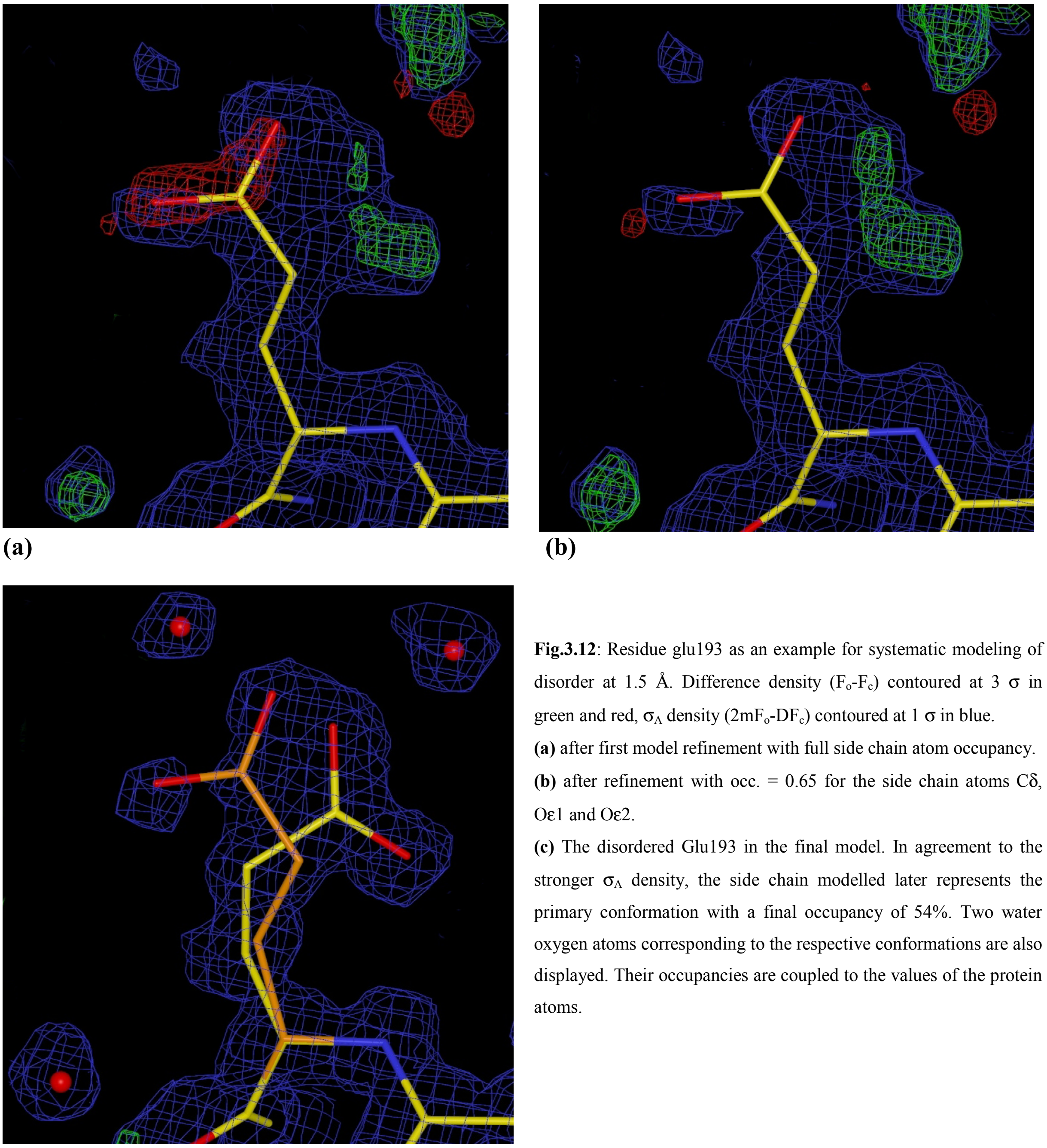
Water oxygen atoms were always refined without including the polar hydrogen atoms of the water molecule. In late stages of the refinement, weaker water peaks were modelled and refined with constant occupancies of 50%. Whenever water networks of half-occupied atoms were overlapping, the corresponding atoms were assigned part numbers *A* or *B*. The same was done for close pairs of partly occupied atoms, representing disordered positions of the same water species. In this case, a common residue number and free occupancy variable was assigned exclusively to the pair. A third type of partly occupied water oxygen atoms are those connected to protein side chain disorder. Atoms with a reasonable distance to the corresponding side chain atom were given the same disorder component (*PART*) number and free occupancy variable.

3.1.7.5 The refinement of conformational disorder

Diffraction data to beyond 2 Å allow the study of detailed model features like side chain disorder. To identify and model atoms with more than one position and non-unit occupancies, the following strategy is successful in most cases. It was applied in all steps of the iterative process of modeling and refinement.

Large difference electron density (F_o-F_c) peaks and holes with a minimum absolute value of 3 σ units in XFIT were further investigated. For amino acid residues where the side chain exhibited significant negative difference density, the occupancy of the atoms was reduced to a fixed value of 65%. In most of these cases, a clear positive difference density was observable after a new refinement at reasonable locations close to the existing side chain and with a reasonable, connected side chain shape. Alternative conformations were then modelled and fitted into the density using the XFIT *split side chain* function. In the following refinement job they were refined as secondary disorder components with a still fixed occupancy of 35%. At later stages in refinement, the occupancies of disordered atoms were refined using free variables in SHELXL with their sum of occupancies constrained to unity. Clusters of disordered residues, connected via hydrogen bonds and water networks, were grouped and refined with common free occupancy variables. These systematic assignments will be explained and discussed in detail in the results chapter.

Well defined twofold disorder like in the case of Glutamate 193 was visible in the early refinement phase (steps 2 – 4) at 1.5 Å. Other residues revealed multiple side chain conformations only later, when higher resolution data were included and after a general improvement of the model. Several problematic cases remained were the weaker or both coformations did not fit the electron density well. Missing σ_A density and remaining positive and negative difference density were observed in such cases. Generally, it was not attempted to refine threefold disorder for the final model. However, threefold disorder was modelled for the two obvious cases of Threonine 244 and Serine 282, and, with less satisfying results, for Methionine 168. Whenever atoms of a new secondary conformation were added to the model in later refinement stages, their displacement parameters were reset to isotropy. This faciliated a more robust SHELXL refinement of new anisotropic B-values.



(a)

(b)

(c)

Fig.3.12: Residue glu193 as an example for systematic modeling of disorder at 1.5 Å. Difference density (F_o-F_c) contoured at 3 σ in green and red, σ_A density ($2mF_o$-DF_c) contoured at 1 σ in blue.

(a) after first model refinement with full side chain atom occupancy.

(b) after refinement with occ. = 0.65 for the side chain atoms Cδ, Oε1 and Oε2.

(c) The disordered Glu193 in the final model. In agreement to the stronger σ_A density, the side chain modelled later represents the primary conformation with a final occupancy of 54%. Two water oxygen atoms corresponding to the respective conformations are also displayed. Their occupancies are coupled to the values of the protein atoms.

3.1.7.6 The placement of hydrogen atoms

The hydrogen atoms were completely ignored until refinement step eight, when the model was in a sufficiently advanced state, including anisotropic displacement parameters, and had been refined against all data. The visibility of difference electron density peaks corresponding to hydrogen atoms was then checked in the F_o-F_c map at a contour level of 2.5 σ (see results). After that, the hydrogen atoms were included into the SHELXL model for the first time. This was done automatically using predefined *HFIX* instructions for all protein, coenzyme and ligand atoms carrying hydrogen atoms. Exceptions were made for all imidazol ring.atoms of histidine, and the tyrosine, serine, threonine and $NADP^+$-phosphate oxygen atoms. All these polar hydrogen atoms as well as water hydrogen atoms were not included into the model, because their visibility in the final difference maps is subject of the structural discussion. During a SHELXL refinement job, the scattering contribution of hydrogen atoms is taken into account by refining them as discrete atoms at geometrically ideal positions (relative to the heavier atoms they are bound to).

3.1.7.7 Including model changes into a new SHELXL instructions file

Model modifications done with XFIT were saved in a new PDB file. Using the SHELXPRO option *update*, the atom lines including all model parameters were transformed from this last PDB file to a new INS file, while all SHELXL-specific instructions and restraints were copied from the last RES file. This operation implies the loss of hydrogen atom lines (*AFIX* lines in the RES files) and free variable number assignments in the atom lines (because the variable numbers, e.g. SHELX code 31.0000, are replaced by the refined variable values, e.g. SHELX code 10.5375, if the value of free variable #3 is 0.5375). Both the *HFIX* instructions to recover the hydrogen atom lines, and the reassignments of free variable numbers were written automatically to the new INS file using a self-made PERL script.

3.1.8 Comparisons of phases and electron density maps

The experimental phases were obtained from SHARP in a binary MTZ format, which was converted to an ascii format PHS file using the CCP4 (Collaborative Computational Project, Number 4, 1994) program MTZ2VARIOUS. Calculated phases from the refinement were produced by SHELXL and written into FCF (text format) files. The refinement phases file used for the phase error determinations and map comparisons was the one obtained after step 11 of the refinement. It will from now on be called the ‘reference’. For the determination of phase differences and map correlation coefficients in reciprocal space, the program SHELXPRO was used. All comparisons were made after shifting the experimental map to the same origin as the reference map. The algorithm used to calculate map correlation coefficients as a function of resolution bins is the one described by Lunin & Woolfson (1993). Besides the reciprocal-space map CC, the figure of merit and the FOM-weighted phase errors, SHELXPRO also determines the cosines of the phase errors and phase errors weighted by both figure of merit and the amplitudes of structure factors. Comparisons were made for all subsequent SHARP maps at different limiting resolutions and for the modified electron density map from SOLOMON. Results are described and discussed in chapter 4.1.2.

Electron density maps were visually compared with XFIT , using a σ_A-weighted map ($2mF_o$-$1DFc$, ϕ_{calc}) for the refined and ($fom*F_o$, ϕ_{MAD}) for the experimental map. Both maps were contoured at 1 σ. In order to measure the real space similarity of the maps quantitatively, the program MAPMAN (Kleywegt & Jones 1996) was used for real-space correlation coefficient calculation. A binary map file in CCP4 format was calculated from the experimental SHARP MTZ file using the CCP4 module FFT. From the refined phases FCF file, another binary map file in DSN6 format was generated using the *M*-option in SHELXPRO. In both transformation processes, identical parameters for map origin, grid spacing and extent were set. Both maps were normalized in MAPMAN (the average density becoming zero and the standard deviation becoming one) before calculating the correlation coefficient.

The CCP4 programs SFALL (Agarwal 1978) and OVERLAPMAP (Branden & Jones 1990) were used to evaluate real-space map correlation coefficients residue per residue. An MTZ-file generated by CCP4/F2MTZ from the experimental PHS file (after the origin shift with SHELXPRO) served as input for the OVERLAPMAP real-space comparison. The corresponding model-calculated map was not stored in any form, but calculated by SFALL from the PDB model *in situ*. The correlation coefficients were determined for each residue, seperately for protein backbone and side chains. Variations in the input PDB model were applied manually to compare only partly occupied side chains of either conformation. Results are presented and discussed in chapter 4.1.3.

3.1.9 The creation of an experimentally phased difference electron density map

F_o and ϕ_{MAD} values from the SHARP-based, origin-shifted PHS file were assigned to the identically indexed structure factors of the SHELXL FCF file, from which only the F_c values were taken. It was checked previously that the amplitudes from both sources were on the same scale. The combined structure factor lines containing the difference density information were written to a new PHS file. All steps mentioned were automatically performed by a self-made Perl script. No reflection-sorting algorithm was needed for the file combination, as the (*h,k,l*) indices were already sorted the same way. Experimental structure factors corresponding to the missing R_{free} set in the refined FCF file were not included in the target PHS file.

The experimentally phased F_o-F_c electron density map was used for the (attempted) bias-free localization of hydrogen atoms.

3.2 The development of the substructure validation program SITCOM

3.2.1 Definition of SITCOM

SITCOM was designed for the comparison of sites resulting from solutions of one or more heavy-atom substructure solving programs. To find equivalent sites with close positions in three-dimensional space, symmetry operators and other space group related features are applied systematically. Sets of sites are scored by the number of sites corresponding to as much other (independent) solutions as possible or to a single reference set of sites from a refined protein model. The positional accuracy, i.e. the mean distance of corresponding site positions is contributing to the score as well.

3.2.2 Program architecture

SITCOM was written in ANSI C. The program functions were grouped into several modular source code files according to functionality. The main program uses functions from a structure managing and comparing core module (*struct.c*), a module organizing input site and crystal cell information (*input.c*) and an output module (*output.c*). The fundamental structure managing functions for transformation and analysis of site positions are based on simpler geometric functions which are responsible for vector and matrix calculations (module *geom.c*). All modules use functions from the most simple module *basics.c*, in which trivial mathematical definitions like for the square function and the output of error messages are defined.

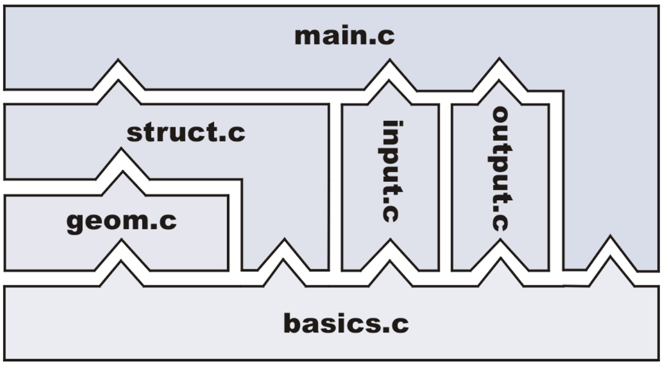


Fig. 3.13 The hierarchical architecture of the SitCom, consisting of functional modules.

3.2.3 Program flow

The flow of operations was designed in a rather linear fashion. The program functions are first concerned with input tasks like text file reading and storage / (re-)organization of sitelists, then with symmetry-related heavy-atom site coordinate modifications and distance analyses, and finally with book-keeping of site correspondancies, with the scoring of sitelists and the creation of output files.

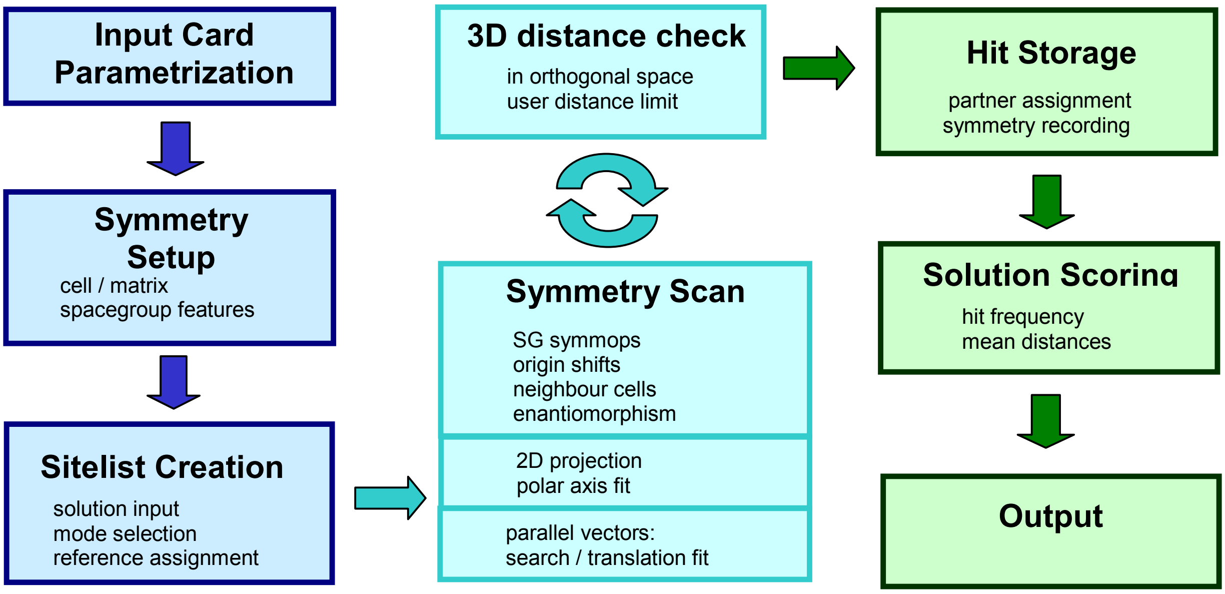
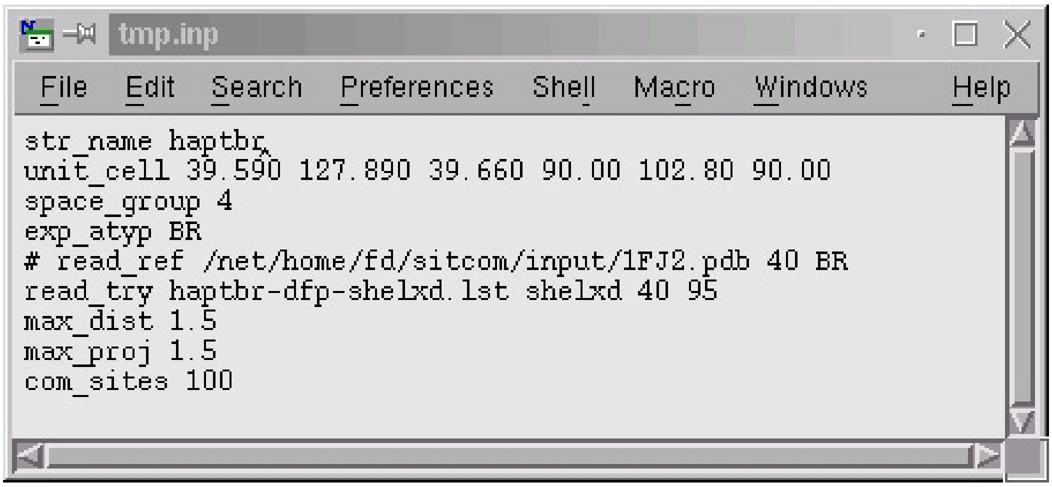


Fig. 3.14: Flowchart for SITCOM. The three columns of boxes represent the major parts of the program flow: Input file-related and preparational tasks are on the left, program functions applying symmetry operations on the sitelists and checking the resulting distances are in the middle column, and result-related scoring and output functions are on the right.

In the first step, SITCOM derives several parameters from input cards that are read from a text file (Fig. 3.15).



```
str_name haptbr
unit_cell 39.590 127.890 39.660 90.00 102.80 90.00
space_group 4
exp_atyp BR
# read_ref /net/home/fd/sitcom/input/1FJ2.pdb 40 BR
read_try haptbr-dfp-shelxd.lst shelxd 40 95
max_dist 1.5
max_proj 1.5
com_sites 100
```

Fig. 3.15 An exemplary input card file. The cards refering to crystal symmetry (**unit_cell** and **space_group**) naturally have constant values for a given structure. At least one **read_try** card is obligatory to define the file location of heavy-atom substructure solutions to be read. A second card for another source file, usually from another solution program, is optional as well as the **read_ref** card for a set of sites extracted from a refined model (PDB format). The last argument of the read_try card is a CC value threshold defining the number of solutions to keep. For example, a value of 95 will cause only solutions with the highest 5% of CC values to be taken. If a set of SHELXD solutions ranges from CC = 50 to 60, only solution trials with CC values greater than 59.5 will be stored. Depending on the distribution of CC values, even a high threshold may keep many solutions. The **max_dist** and **max_proj** cards define the distance limits for sites to be regarded as equivalent, the first is applied in three dimension, the second only for a two-dimensional projection. The **com_sites** card defines in how many solutions a site has to be found in order to be kept, e.g. in all solutions if the parameter is 100 (%).

All crystallographic information needed by SITCOM for a given space group number is extracted from a symmetry library, being a text file containing each operator as separate line.

```
1 P1            1  4  1  1  1        x        y      z
2 P-1           0  0  2  2  2        x        y      z
2 P-1           0  0  2  2  2       -x       -y     -z
3 P2            1  2  2  1  2        x        y      z
3 P2            1  2  2  1  2       -x        y     -z
4 P2(1)         1  2  2  2  2        x        y      z
4 P2(1)         1  2  2  2  2       -x    y+1/2     -z
5 C2            1  2  2  2  2        x        y      z
5 C2            1  2  2  2  2       -x        y     -z
5 C2            1  2  2  2  2    x+1/2    y+1/2      z
5 C2            1  2  2  2  2   -x+1/2    y+1/2     -z
```

Fig. 3.16: Part of the symmetry library used by SITCOM. Each single operator is written in a separate line. The colums are (a) space group number , (b) space group symbolic description, (c) acentricity flag (1 = yes), (d) type of polarity / floating origin axis (0 = apolar, 1, 2, 3 = polar axis is a, b, c, 4 = special case P1 with three floating origin directions), (e, f, g) the number of origins along each cell axis, (h, i, j) the three fractional coordinate transformators of the symmetry operator. This text file was created with a PERL program from the symmetry library of CNS (Brünger et *al.* 1998). The relevant symmetry operators are read from the file in the beginning of a SITCOM job.

Having stored all crystallographically relevant information, the substructure files are read as defined by the respective cards. Each sitelist is stored in an internal program format and marked by an identification tag, which is related to the program source (SHELXD, SOLVE…), and the solution number – *e.g. shelxd.11*. Also the CC(E_{obs}, E_{calc}) values (see theory) are used – they are not important for the final SitCom solution score, but neccessary for the initial sitelist selection.

The stored properties of single sites are the site number, the fractional/ orthogonal coordinates and the peak height. The rectangular coordinates are calculated by SITCOM. In case of a PDB-derived sitelist (the so-called reference set) reciprocal B-values, scaled to a maximum of 99.99, are used as pseudo peak heights.

After all sitelists have been organized, SITCOM starts the symmetry scan for each combination of working and reference set of sites. The principles of substructure comparison, as explained in the following section, are the same for (a) only two substructures, (b) many substructures compared to a single reference set and (c) the cross-comparison of many substructures against many reference sets. The comparison mode depends on the input settings: If the user has supplied a refined PDB substructure, the solution trials are only compared to this reference set. Otherwise, all solutions from one or more file source are cross-compared to each other, sequentially using every substructure as "pseudo-reference".

The three-dimensional analysis of distances between sites for a given combination of working and reference substructure is done repeatedly for every symmetry equivalent of the working set. The number of sites close to a equivalent reference set partner and the mean distance between all pairs of sites are used to score a substructure. Details about the scoring process are explained in section 3.2.7. The trial solution with the highest score is written to several output files such *output.res* (SHELX format), *output.hatom* (SHARP format), *output.pdb* (CCP4 format).

3.2.4 General algorithms for site comparison

Most of the SITCOM routines are sequences of basic algebraic operations like the multiplication of a vector with a matrix. The corresponding program function is used to apply symmetry operators and to transform fractional site coordinates into orthogonal ones and vice versa – the necessary transformation matrix elements are derived from the symmetry library or from cell constants, respectively. All symmetry operations are applied in fractional space. The calculation of distances between the sites of the given two compared solutions is applied in orthogonal space, cycling over both sets of sites in two nested loops, i.e. comparing all possible combinations $S_{i,j}$ of sites.

The main symmetry scan routine applies an outer loop over all given space group symmetry operators, and inner loops over the origin shifts. The *3D* distance check is then applied for every symmetry-equivalent image of a working sitelist. For instance, four symmetry operators exist in spacegroup $P2_12_12_1$. There are two origin setups per cell axis – including the neighbour cells, which are scanned as well, so seven fixed origin shifts have to be applied (at fractional positions -1, -0.5, 0, 0.5, 1, 1.5, 2). This leads to 4 x 7 x 7 x 7 = 1372 sitelist images, which are sequentially generated and compared to the given reference. The whole procedure is also applied to the inverted coordinates of the probe sites, so that 2744 comparison operations per sitelist combination result for the $P2_12_12_1$ example. During the symmetry scan process, every positional agreement between working set and reference set sites is recorded. Double assignments are eliminated by selecting the pairs with shorter respective distances. Finally, the number of the corresponding reference site partner as well as the distance are stored for the scoring.

This general procedure is applied for all structures in non-polar space groups. The basic principle is the same for the special cases of polar space groups and P1. Special algorithms will be explained in the following.

3.2.5 Polar spacegroups

SitCom recognizes polar spacegroups and the direction of polar axes based on the information in the symmetry library. In case of a polar spacegroup, the sites are treated in a different way than for non-polar spacegroups. The algorithm consists of two parts, the first being a variation of the normal symmetry loop. The symmetry operators are applied as usual, but in the inner loops, only origin shifts in the plane perpendicular to the polar axis are applied. For each site combination of two solutions to be compared, a distance check is done in this non-polar projection plane only. Possibly corresponding sites are stored in a temporary list as preliminary projection pairs.

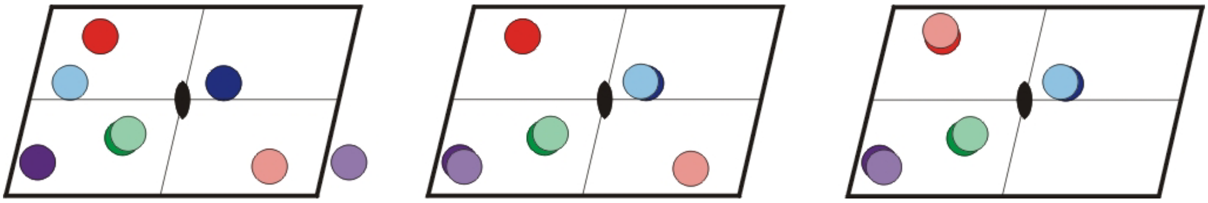


Fig. 3.17: The projectional symmetry loop algorithm for a polar spacegroup (P2_1). The principle is the same for non-polar spacegroups, but all operations are then applied in three dimensions. **Left**:Four pairs of corresponding sites coloured differently – the lightly coloured circles symbolize the trial sites, the dark ones the reference sites. The green pair is already in similar positions. **Middle:** Applying an origin shift of ½ in c for the blue site and a neighbour cell shift of –1 in c for the purple site leads to positional agreement. **Right:** For the red site, the symmetry operator of –x, y+½, -z is applied. The shift of ½ in b can be neglected as it is perpendicular to the projection plane. **Note:** In this scheme, the origin shift of ½ in c is applied to one single site only to explain the principle. In reality, SitCom applies the same origin shifts to all sites of a solution (whereas symmetry operators may differ between sites of the same list).

If all possible pairs of sites are found, their relation along the polar axis is analyzed and refined. The mean shift along the axis is calculated for the ensemble of pairs after correcting neighbour cell displacements of single sites (in the polar direction). Pairs differing significantly from the mean shift are deleted from the preliminary list of equivalents. The mean displacement is refined by iteratively repeating the procedure with decreasing tolerances for mean shift deviations. The pairs remaining after the last of five refinement and correction cycles finally undergo the regular three-dimensional distance check as for non-polar spacegroups.

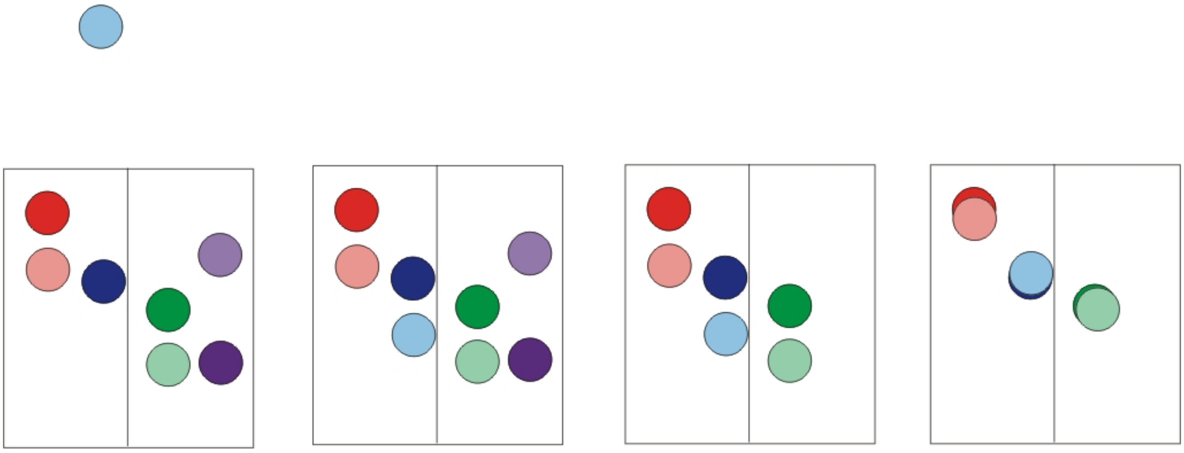


Fig. 3.18: The polar axis shift algorithm for a polar spacegroup (P21), explained for the pairs of sites as assigned after the projection plane fit (Fig. 3.17, now using the the same colours). (a): For the red and green pair the distance along b is already comparable. The bright blue test site is additionally displaced by one cell edge length. The purple pair has a totally different distance. (b): After applying the neighbour cell correction for the blue site, its distance to the reference has become similar. (c) SitCom has calculated the mean shift of all four pairs and found the purple pair to be deviating too much from it, therefore this pair has been discarded. (d) For the remaining three pairs, the (non-absolute) sum of deviations from the average shift has become zero. The shift is applied to the trial sites.

3.2.6 The P1 algorithm

To determine the common arbitrary shift between any two substructures compared, as present in spacegroup P1 because of three freely chosen origins, the following algorithm is applied.

The first part of the procedure is performed on the fractional site coordinates. SitCom tries to find two parallel translation vectors for two pairs of heavy atom sites, pairing every site from one list with every site from the other one, also checking for whole-cell-edge displacements. As soon as two identical vectors (within a tolerance) are found, the averaged translation is applied to all sites of the probe substructure.

In the second phase, a distance check in orthogonal space is applied. For sites of both sets that remain without partner afterwards, a cell adjustment is tested – one-directional cell-edge shifts are applied to the test sites, if their fractional distances to any reference position become smaller. Having adjusted sites, the orthogonal distance check routine is repeated. The two processes of adjusting and checking are repeated 10 times at most, or abandoned if all probe sites have been matched with a different partner.

The third phase of the P1-handling algorithm is an orthogonal-space refinement of the initial *3D*-superposition, in order to minimize the mean distance of corresponding sites. Towards this a three-dimensional grid search is carried out. To save computing time, the test set sites do not undergo the grid search individually with subsequent mean distance calculation. Instead, the centers of mass are determined initially for both substructures, the one for the working set already including the approximate translation. A box of one cubic Ångstrom around the translation position is scanned in

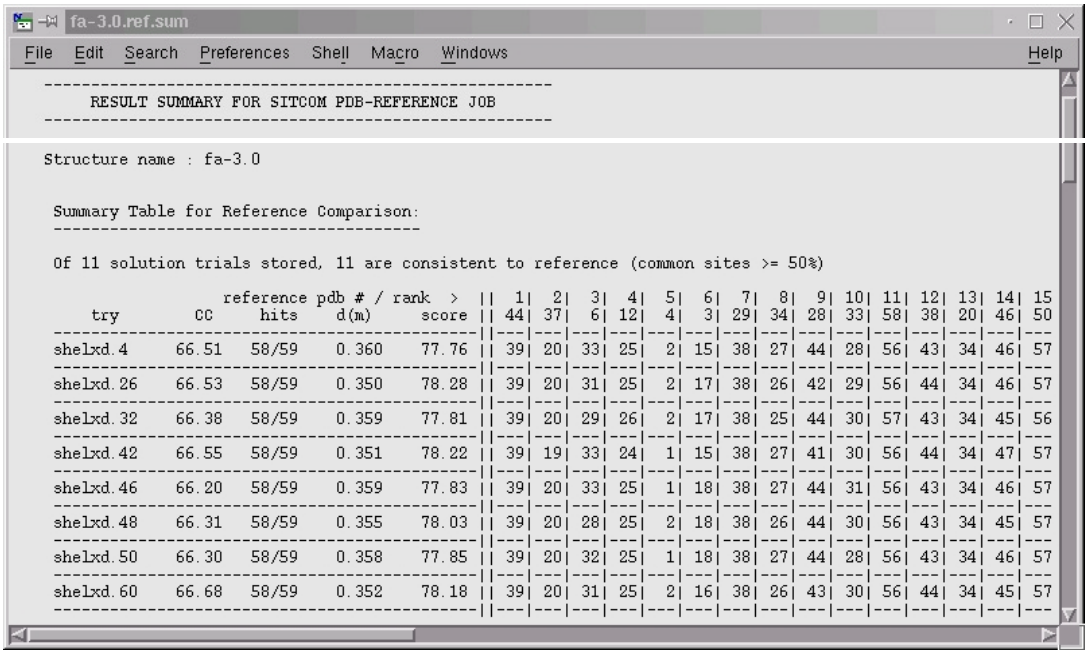
0.1 Å steps. The shift-correction causing the centers of mass to be closest is added to the shift derived in phase two. The refined translation is finally applied to the sites of the working solution, and the usual distance check and partner assignment is done as for all spacegroups.

3.2.7 Substructure scoring

Once all symmetry transformations, distance checks, and the bookkeeping of corresponding sites have been done, SITCOM evaluates the results using special functions. The score for a trial substructure is based (a) on the number of sites identified as equivalent, relative to the total number of reference positions and (b) on the mean distance between these positionally agreeing sites:

$$\text{score [\%]} = 100 * \sqrt{\left(\frac{n_{try}}{n_{ref}}\right)^2 * \frac{1}{\left(\exp\left(\bar{d}\right)\right)^2}}$$

In case of the *reference comparison mode*, the final scores of the solution trials are calculated directly.

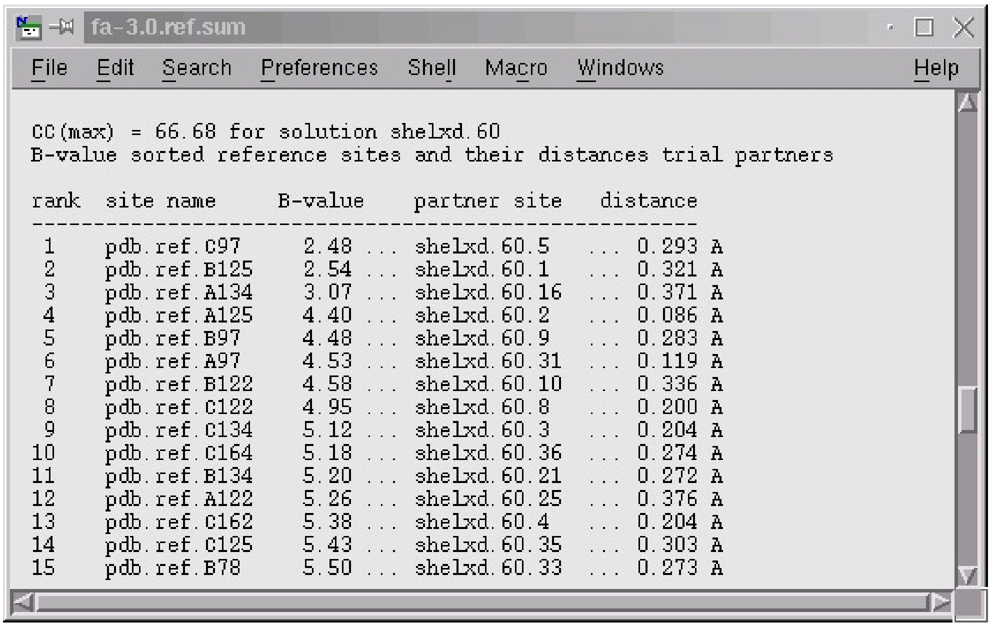


```
    ---------------------------------------------------
        RESULT SUMMARY FOR SITCOM PDB-REFERENCE JOB
    ---------------------------------------------------

    Structure name : fa-3.0

     Summary Table for Reference Comparison:
     ---------------------------------------

     Of 11 solution trials stored, 11 are consistent to reference (common sites >= 50%)

                          reference pdb # / rank  > ||  1|  2|  3|  4|  5|  6|  7|  8|  9| 10| 11| 12| 13| 14| 15
         try        CC     hits     d(m)     score || 44| 37|  6| 12|  4|  3| 29| 34| 28| 33| 58| 38| 20| 46| 50
     -----------------------------------------------||---|---|---|---|---|---|---|---|---|---|---|---|---|---|---
     shelxd.4      66.51   58/59    0.360     77.76 || 39| 20| 33| 25|  2| 15| 38| 27| 44| 28| 56| 43| 34| 46| 57
     -----------------------------------------------||---|---|---|---|---|---|---|---|---|---|---|---|---|---|---
     shelxd.26     66.53   58/59    0.350     78.28 || 39| 20| 31| 25|  2| 17| 38| 26| 42| 29| 56| 44| 34| 46| 57
     -----------------------------------------------||---|---|---|---|---|---|---|---|---|---|---|---|---|---|---
     shelxd.32     66.38   58/59    0.359     77.81 || 39| 20| 29| 26|  2| 17| 38| 25| 44| 30| 57| 43| 34| 45| 56
     -----------------------------------------------||---|---|---|---|---|---|---|---|---|---|---|---|---|---|---
     shelxd.42     66.55   58/59    0.351     78.22 || 39| 19| 33| 24|  1| 15| 38| 27| 41| 30| 56| 44| 34| 47| 57
     -----------------------------------------------||---|---|---|---|---|---|---|---|---|---|---|---|---|---|---
     shelxd.46     66.20   58/59    0.359     77.83 || 39| 20| 33| 25|  1| 18| 38| 27| 44| 31| 56| 43| 34| 46| 57
     -----------------------------------------------||---|---|---|---|---|---|---|---|---|---|---|---|---|---|---
     shelxd.48     66.31   58/59    0.355     78.03 || 39| 20| 28| 25|  2| 18| 38| 26| 44| 30| 56| 43| 34| 45| 57
     -----------------------------------------------||---|---|---|---|---|---|---|---|---|---|---|---|---|---|---
     shelxd.50     66.30   58/59    0.358     77.85 || 39| 20| 32| 25|  1| 18| 38| 27| 44| 28| 56| 43| 34| 46| 57
     -----------------------------------------------||---|---|---|---|---|---|---|---|---|---|---|---|---|---|---
     shelxd.60     66.68   58/59    0.352     78.18 || 39| 20| 31| 25|  2| 16| 38| 26| 43| 30| 56| 44| 34| 45| 57
     -----------------------------------------------||---|---|---|---|---|---|---|---|---|---|---|---|---|---|---
```

Fig. 3.19: An extract of the SITCOM summary file for a comparison between SHELXD substructure solution trials and refined selenium positions from a single reference PDB file. This example is for the Transhydrogenase structure, see chapter 4.2.2.

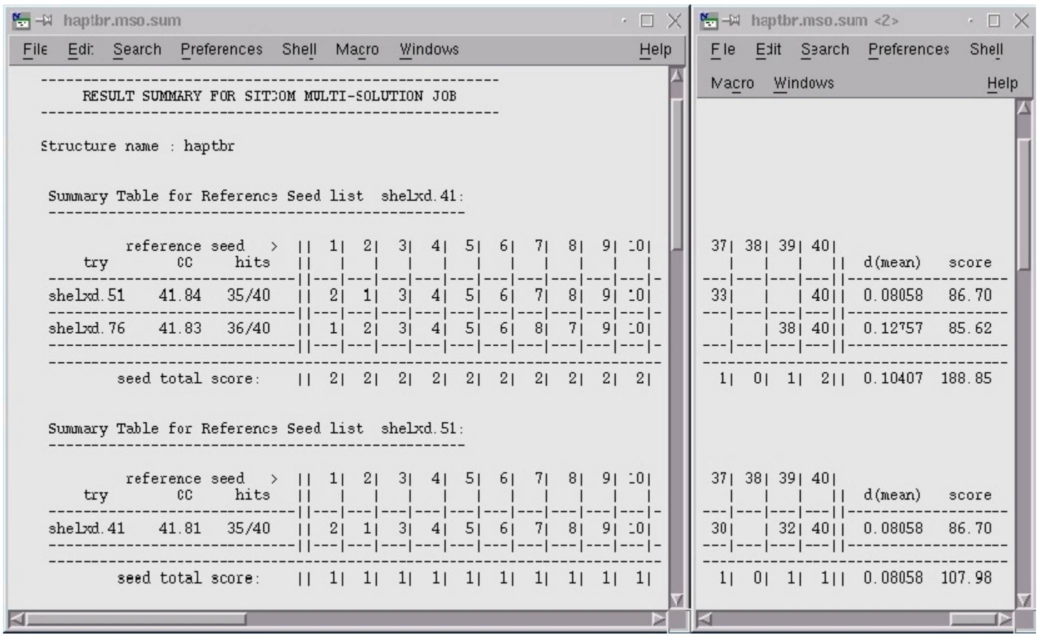
In the summary file, a list of all successful solution trials is given, where “successful” is defined by a correspondancy rate of at least 50% to the reference site set. Besides the hit rate, the mean distance and the resulting score, a detailed table of site relationships is listed. The columns refer to the heavy atom numbers in the PDB file, and each field of the trial solution lines contains the number of the corresponding probe site. Further information in the summary file includes a table focussing on the

PDB file heavy atoms (Fig. 3.20) in order to show the correlation of refined B-values to trial site peak heights and pair distances, as discussed in chapter 4.2.2.



```
CC(max) = 66.68 for solution shelxd.60
B-value sorted reference sites and their distances trial partners

rank  site name     B-value    partner site   distance
-------------------------------------------------------
 1    pdb.ref.C97     2.48 ... shelxd.60.5    ... 0.293 A
 2    pdb.ref.B125    2.54 ... shelxd.60.1    ... 0.321 A
 3    pdb.ref.A134    3.07 ... shelxd.60.16   ... 0.371 A
 4    pdb.ref.A125    4.40 ... shelxd.60.2    ... 0.086 A
 5    pdb.ref.B97     4.48 ... shelxd.60.9    ... 0.283 A
 6    pdb.ref.A97     4.53 ... shelxd.60.31   ... 0.119 A
 7    pdb.ref.B122    4.58 ... shelxd.60.10   ... 0.336 A
 8    pdb.ref.C122    4.95 ... shelxd.60.8    ... 0.200 A
 9    pdb.ref.C134    5.12 ... shelxd.60.3    ... 0.204 A
10    pdb.ref.C164    5.18 ... shelxd.60.36   ... 0.274 A
11    pdb.ref.B134    5.20 ... shelxd.60.21   ... 0.272 A
12    pdb.ref.A122    5.26 ... shelxd.60.25   ... 0.376 A
13    pdb.ref.C162    5.38 ... shelxd.60.4    ... 0.204 A
14    pdb.ref.C125    5.43 ... shelxd.60.35   ... 0.303 A
15    pdb.ref.B78     5.50 ... shelxd.60.33   ... 0.273 A
```

Fig. 3.20: Extract from the same summary file as for the previous figure, displaying the PDB heavy atom statistics.


haptbr.mso.sum
File Edit Search Preferences Shell Macro Windows Help
haptbr.mso.sum <2>
File Edit Search Preferences Shell
Macro Windows Help

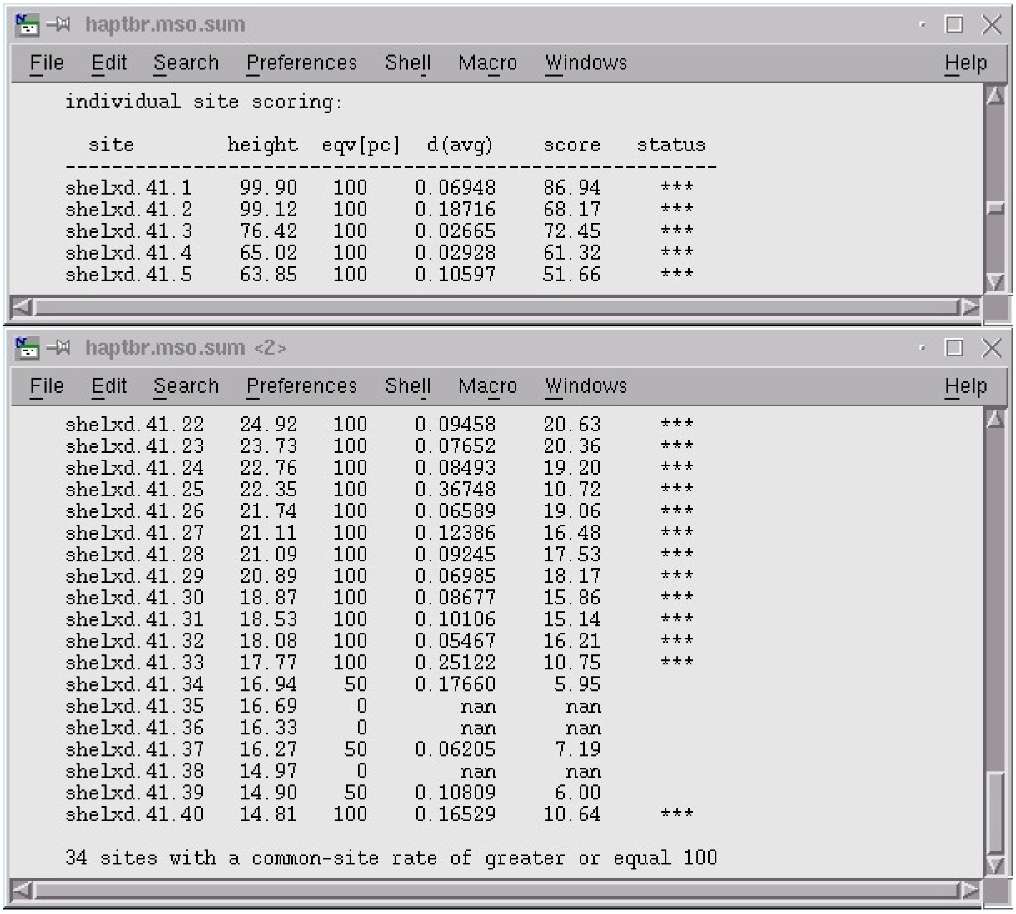
```
 ------------------------------------------------------
      RESULT SUMMARY FOR SITCOM MULTI-SOLUTION JOB
 ------------------------------------------------------

 Structure name : haptbr

  Summary Table for Reference Seed list  shelxd.41:
  ------------------------------------------------

           reference seed   >  ||  1|  2|  3|  4|  5|  6|  7|  8|  9| 10|
     try         CC     hits   ||   |   |   |   |   |   |   |   |   |   |
  -----------------------------||---|---|---|---|---|---|---|---|---|---|-
  shelxd.51    41.84    35/40  ||  2|  1|  3|  4|  5|  6|  7|  8|  9| 10|
  -----------------------------||---|---|---|---|---|---|---|---|---|---|-
  shelxd.76    41.83    36/40  ||  1|  2|  3|  4|  5|  6|  8|  7|  9| 10|
  -----------------------------||---|---|---|---|---|---|---|---|---|---|-
  ------------------------------------------------------------------------
          seed total score:    ||  2|  2|  2|  2|  2|  2|  2|  2|  2|  2|

  Summary Table for Reference Seed list  shelxd.51:
  ------------------------------------------------

           reference seed   >  ||  1|  2|  3|  4|  5|  6|  7|  8|  9| 10|
     try         CC     hits   ||   |   |   |   |   |   |   |   |   |   |
  -----------------------------||---|---|---|---|---|---|---|---|---|---|-
  shelxd.41    41.81    35/40  ||  2|  1|  3|  4|  5|  6|  7|  8|  9| 10|
  -----------------------------||---|---|---|---|---|---|---|---|---|---|-
  ------------------------------------------------------------------------
          seed total score:    ||  1|  1|  1|  1|  1|  1|  1|  1|  1|  1|
```

```
 37| 38| 39| 40|
   |   |   |   || d(mean)   score
---|---|---|---||-------------------
 33|   |   | 40|| 0.08058   86.70
---|---|---|---||-------------------
   |   | 38| 40|| 0.12757   85.62
---|---|---|---||-------------------
-----------------------------------
  1|  0|  1|  2|| 0.10407  188.85


 37| 38| 39| 40|
   |   |   |   || d(mean)   score
---|---|---|---||-------------------
 30|   | 32| 40|| 0.08058   86.70
---|---|---|---||-------------------
-----------------------------------
  1|  0|  1|  1|| 0.08058  107.98
```

Fig. 3.21: Extract of a summary file for multi-solution cross-comparison of the HAPTBr Bromine substructures (see chapter 4.2.3). Only three solutions – shelxd.41,51,76 – were compared due to an initial 90% CC value threshold. Solution shelxd.41 agrees to 51 and 76, but shelxd.51 not to 76. Therefore, shelxd.41 has a double-weighted score relative to shelxd.51, which is much higher despite the higher overall mean site distance (score values are 189 and 108, respectively – see rightmost column).

Finally, a histogram for trial substructure hit rates is given. This is useful to examine the quality of the whole compared solution set, for example consisting of 100 trials. A closer discussion, concerned with the substructure accuracy obtained from different wavelength data subsets, is given in chapter 4.2.2.

If SitCom is in multi-solution cross-comparison mode, the substructures are scored by their consistency to all other solutions. Therefore, the final score is derived from the individual trial substructure scores that have been determined before. The part of the summary file showing the site relationships is the same as for the reference comparison mode, but here, for every pseudo-reference list, a separate table is printed. The sub-scores of the tested trial site lists, displayed in the rightmost column of Fig. 3.21, are calculated the usual way from the hit rates and the mean distances. The final reference-solution score of interest is calculated from the mean score of all consistent solutions (with the same hit rate > 50% criterium as before), multiplied with a factor taking the number of agreeing solutions into account.



```
individual site scoring:

  site        height  eqv[pc]  d(avg)     score   status
---------------------------------------------------------
shelxd.41.1     99.90   100    0.06948    86.94     ***
shelxd.41.2     99.12   100    0.18716    68.17     ***
shelxd.41.3     76.42   100    0.02665    72.45     ***
shelxd.41.4     65.02   100    0.02928    61.32     ***
shelxd.41.5     63.85   100    0.10597    51.66     ***
```

```
shelxd.41.22    24.92   100    0.09458    20.63     ***
shelxd.41.23    23.73   100    0.07652    20.36     ***
shelxd.41.24    22.76   100    0.08493    19.20     ***
shelxd.41.25    22.35   100    0.36748    10.72     ***
shelxd.41.26    21.74   100    0.06589    19.06     ***
shelxd.41.27    21.11   100    0.12386    16.48     ***
shelxd.41.28    21.09   100    0.09245    17.53     ***
shelxd.41.29    20.89   100    0.06985    18.17     ***
shelxd.41.30    18.87   100    0.08677    15.86     ***
shelxd.41.31    18.53   100    0.10106    15.14     ***
shelxd.41.32    18.08   100    0.05467    16.21     ***
shelxd.41.33    17.77   100    0.25122    10.75     ***
shelxd.41.34    16.94    50    0.17660     5.95
shelxd.41.35    16.69     0        nan      nan
shelxd.41.36    16.33     0        nan      nan
shelxd.41.37    16.27    50    0.06205     7.19
shelxd.41.38    14.97     0        nan      nan
shelxd.41.39    14.90    50    0.10809     6.00
shelxd.41.40    14.81   100    0.16529    10.64     ***

34 sites with a common-site rate of greater or equal 100
```

Fig. 3.22 Extract of the site selecting part of the SitCom summary file for the best solution. Only sites with 100% agreement to other solutions are selected (flagged ***).

In cases, where the original CC(E_{obs},E_{calc}) values of solutions and eventually also the SitCom scores are very similar (which is a possible situation for solutions from only one program) a more important question would be, which sites of a given solution are reliable. Therefore, the columns of a cross-comparison table (Fig. 3.21) are evaluated, counting the number of agreements to other substructures for a given reference-solution site. The user initially defines the percentage of site agreements to other solutions serving as a selection threshold. SitCom prints a separate list for the highest-scored solution and flags all sites that have passed the agreement test, thus recommending those sites as reliable for use in phasing.

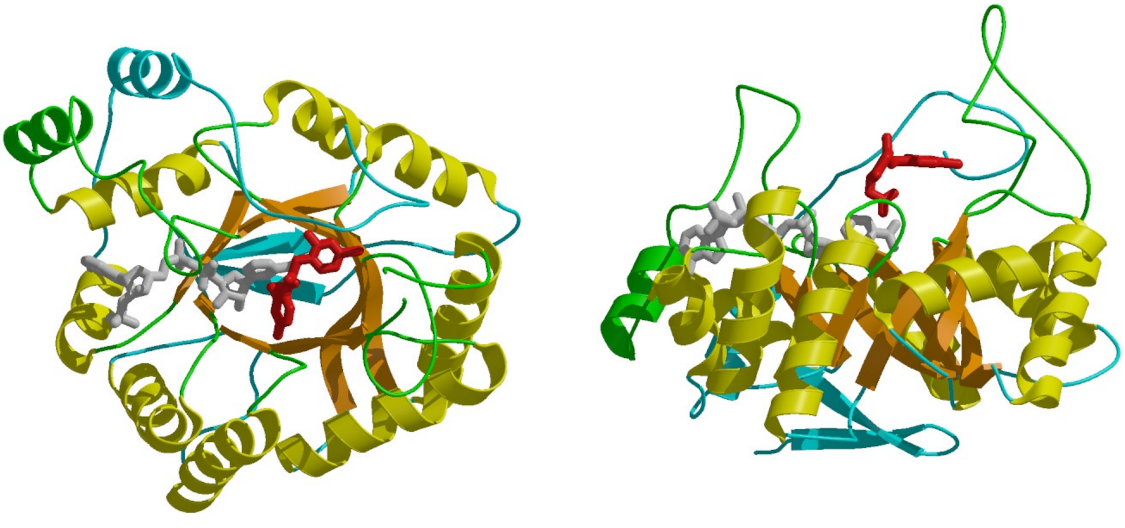
4 Results and Discussion

4.1 Results for the crystal structure of human Aldose Reductase

4.1.1 Refinement characteristics and description of the refined hAR2 model

4.1.1.1 Structural overview

Human Aldose Reductase shows a typical $(\beta\alpha)_8$ TIM-barrel fold, with additional secondary structure elements not being part of the barrel motif: A N-terminal β-sheet consisting of two strands, located at the bottom of the barrel fold and oriented perpendicular to the barrel, a C-terminal helix and a second helix placed between strand and helix of the seventh barrel repeat. The two non-barrel helices are neighbouring each other parallely and are located at the outside of the barrel fold. Ligand and coenzyme are found on top of the barrel motif. The $NADP^+$ is buried more deeply into the protein, close to the barrel elements, whereas the ligand is surrounded by the loop region. Thus, the ligand molecule connects several loops and stabilizes a positionally fixed form of the tertiary structure, which obviously is an important reason for the extraordinarily good diffraction of the hAR crystals.



(a) **(b)**

Fig. 4.1: The tertiary structure of human Aldose Reductase, presented as protein cartoon. **(a)** view on top of the TIM barrel, **(b)** view on its side. The barrel sheet is coloured orange, the barrel helices yellow. The additional elements are coloured cyan and green, the ligand red and the coenzyme grey.

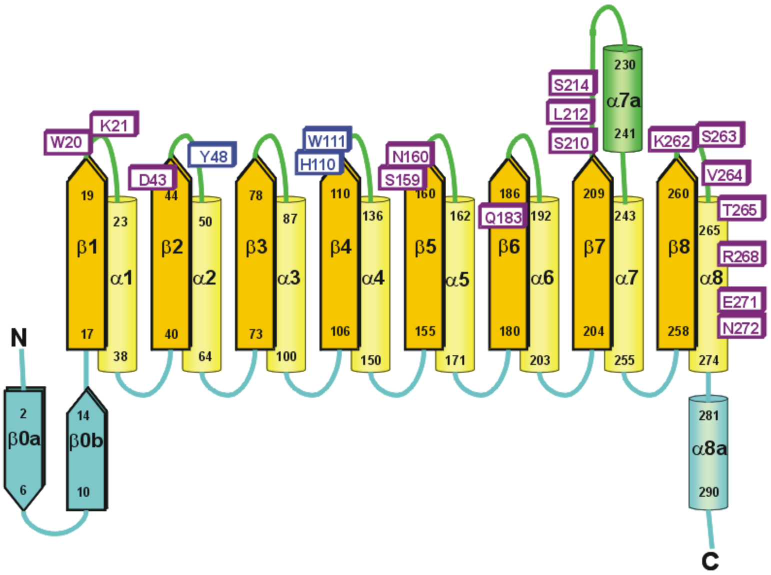


Fig. 4.2: Topology diagram for the tertiary structure of hAR2. Residues interacting with the coenzyme NADP$^+$ (purple) and with the ligand (blue) are explicitly named in the boxes.

The topology diagram reveals that residues of most of the barrel modules (16 residues), especially of helix α_8, are interacting with the coenzyme, but only three residues belonging to two loops show interaction with the ligand.

4.1.1.2 Quality of the refined model

Refining the human Aldose Reductase model with SHELXL lead to a final R_{work} of 8.0% and a R_{free} of 9.6%, respectively. For 71 of the 313 modelled amino acid residues, double conformations, reduced occupancies or atom type disorder were found. The final model contains 842 waters, of which 416 are connected to protein disorder, are disordered in double positions or simply partly occupied. Two citrate molecules from the crystallization buffer were found, both with reduced occupancy.

The Ramachandran diagram (Fig. 4.3) indicates that the secondary structure has a good agreement to expected torsion angle values: 89.8% of the residues are in the most favoured region for torsion angles, the rest is in the additionally allowed region. The B-values for the protein are relatively low (6.6 and 9.7 Å^2 for the backbone and side chains respectively) and show an expected correlation to the protein topology with higher B-values on the surface. The C-terminus however has a much less localized electron density distribution than the rest of the protein. Therefore the last three residues 313-315 (having an average B-value of 27.9 Å^2) were removed from the model.

Global Parameters	
resolution used in final refinenement [Å]	19.3-0.90
data	209774
parameters	32004
restraints	41024
R_{work} [%]	8.0
R_{free} [%]	9.6
Average B-values [$Å^2$]	
main chain atoms	6.6
side chain atoms	9.7
ligand	5.1
coenzyme	4.6
waters	27.2
Ramachandran Statistics [%]	
residues in most favoured regions	89.8
residues in additional allowed regions	10.2

Table 4.1: Overview of general refinement results

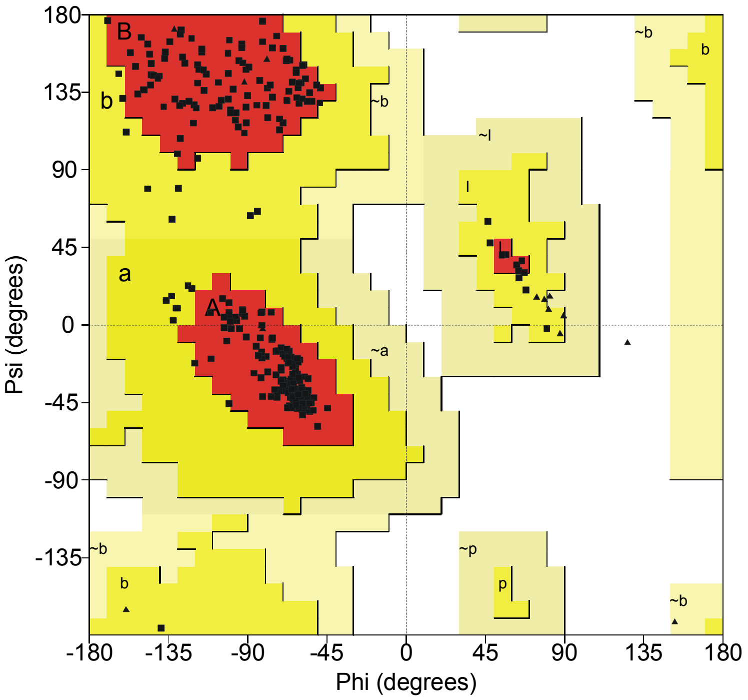


Fig. 4.3: Ramachandran diagram of dihedral angles for the mainchain of the hAR2 model. (Ramachandran & Sassiekharan 1968)

Disagreements between restaint target values and observed model parameter values can be valuable to detect misplaced side chains or other errors of the refinement. Most of the restraint violations found after the last SHELXL cycle of the final hAR2 refinement step are listed in the appendix (Tables A).

The highest deviation in total is 16.8 σ for the planarity restraint (FLAT, Table A.1) of the phenyl ring of residue Tyr209. Comparing the model to the experimental MAD map, the not strictly planar model geometry fits the observation very well. An attactive π-stacking interaction with the almost coplanar NADP$^+$ nicotinamide ring may explain this. The result underlines the fact that data of high quality and resolution (with a good data-to-parameter ratio) can dominate restraints in refinement.

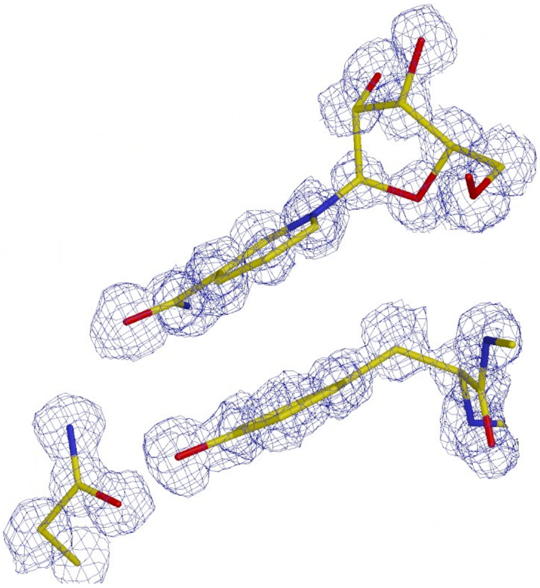


Fig. 4.4: Geometry of residue Tyr209, displayed with the experimental map, (F_{obs}, ϕ_{MAD}) at 1σ contour.

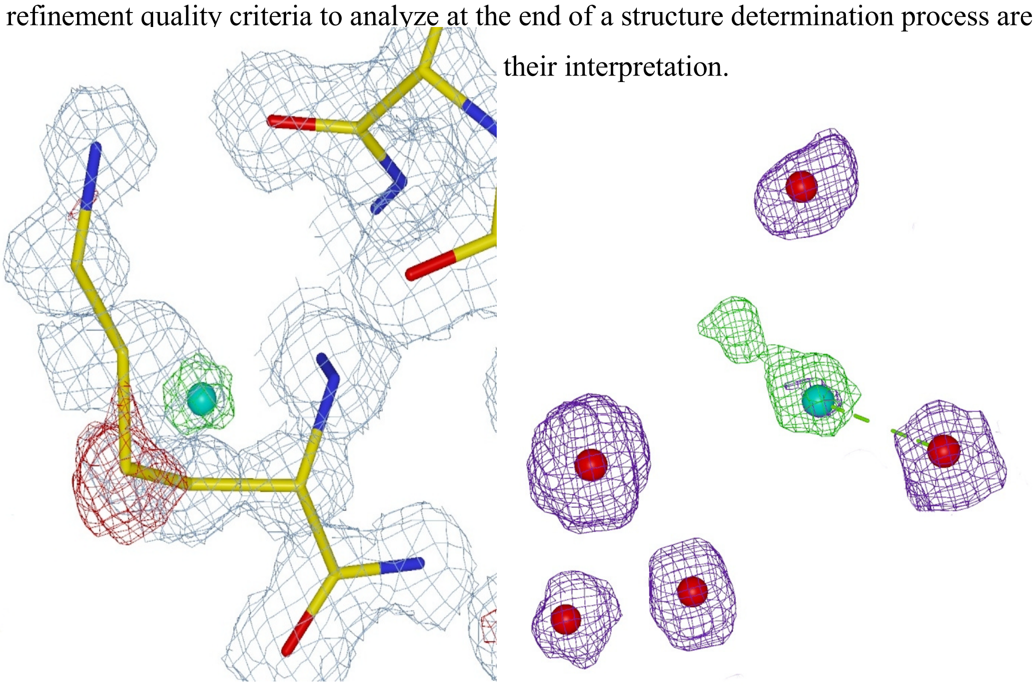
The phenyl ring of Tyr209 is slightly bent towards the nicotinamide ring of the coenzyme (Fig. 4.4) explaining both the FLAT deviation and a disagreement of about 5.5 σ to the anti-bumping restraint for the atoms $C\delta_1$ of the tyrosine and atom C17 of NADP$^+$(Table A.3). Many of the other larger BUMP violations, including the largest one, are related to close contacts between side chain carbon atoms and the backbone (carbonyl) carbon. The observed value of about 3.2 Å is found in many protein structures, and it seems that the BUMP target value of 3.3 Å is too strict in this case.

The agreement of observed values to target values is satisfactory for all parameters related to displacement restraints (Tables A.4-A.6), the B values deviate less than 5 σ from their targets. The atoms violating the SIMU restraints (Table A.4), *i.e.* showing significant ADP deviations to neighbouring atoms, are all belonging to disordered protein side chains or waters which have reduced occupancies.

The deviations concerning distance restraints are higher, but still acceptable. Both the bond length and angle restraints have the largest disagreements around 7 σ. It is noteable that ten of the eleven largest DFIX violations (Table A.7) represent bonds being shorter than the expected distance.

Other refinement quality criteria to analyze at the end of a structure determination process are the

their interpretation.



(a) **(b)**

Fig. 4.5 The two highest remaining difference electron density peaks, **(a)** peak one located at residue Lys100, **(b)** peak two at 1.77 Å distance to the existing water oxygen atom 3117A.

Peak	abs. Height	sigma units	location/ interpretation
1	0.66	9.57	Lys100 – alternative conformation cannot be modelled
2	0.64	9.28	secondary water position (type III) close to HOH 3117A
3	0.59	8.55	secondary water position (type III) close to HOH 3004A
4	0.51	7.39	B-part water (type IV) close to Asp297A
5	0.51	7.39	A-part water (type IV) close to Ser305B
6	0.50	7.25	doubtful peak without σ_A density
7	0.50	7.25	doubtful peak without σ_A density
8	0.49	7.10	weak peak at tertiary Met168C conformation
9	0.48	6.96	doubtful peak without σ_A density
10	0.48	6.96	doubtful peak without σ_A density

Table 4.2: the remaining difference electron density peaks (absolute height given in e/Å^3) after finishing the hAR2 refinement process. One sigma is 0.069 e/Å^3. For the definition of water types, see Table 4.4.

The highest peak is located very close to the Cγ atom position of residue Lys100, in agreement to a negative difference density minimum of –6 σ at the Cγ atom position itself. While this observation clearly indicates that at least one different secondary conformation must exist for Lys100, it is not possible to model an alternative position for the side chain without changing the whole residue location including the backbone. Furthermore, there are no other difference density maxima or minima present which would underpin a second conformation or even a lower occupancy of the side chain modelled. Therefore this peak is regarded as a not sufficiently explainable.

4.1.1.3 Features of intermediate refinement states for different data at 1.5 Å

Before proceeding with the refinement against the high energy remote data, the first four iterative modeling and refinement steps were analogously applied for the other two data subsets. Therefore three intermediate structures of hAR2 at 1.5 Å resolution were present, each based on an isotropic displacement model, and containing most of the fully occupied water molecules plus about half of the clearly visible double side chain conformations.

Data subset	high-en. remote	peak	inflection point
data properties at 1.5 Å			
number of reflections	49055	49105	48719
% complete	99.4	99.5	98.7
I/σ(I)	33.4	28.1	28.5
R_{int}	3.71	5.44	3.52
R_{sigma}	2.70	3.28	2.98
refinement statistics			
number of working reflections	46559	46613	46248
number of free reflections	2496	2492	2471
parameters	12091	12059	12117
restraints	11081	11017	11112
R_{work}	0.1623	0.1542	0.1632
R_{free}	0.1969	0.1934	0.2008
model characteristics			
number of water molecules	343	345	346
number of disordered side chains	25	22	25
max. difference electron density peak	2.33	1.73	2.04
mean B-value for protein main chain	7.76	7.14	6.70
mean B value for protein side chains	11.43	11.21	11.21
mean B-value for water oxygen atoms	25.99	25.28	26.32

Table 4.3: comparative overview of the general 1.5 Å refinement results against the three MAD-wavelength data subsets.

The results for the three data sets are quite similar. According to most of the quality indicators, the refinement against the peak data intensities leads to somewhat better results than the other two refinements. It is still an open question whether this would generally favour the selection of peak data for structure refinement.

After presenting and discussing the formal state and quality of model and refinement at the end of the hAR2 structure determination, features of the structure – protein, ligands and solvent water molecules – shall be explained and discussed.

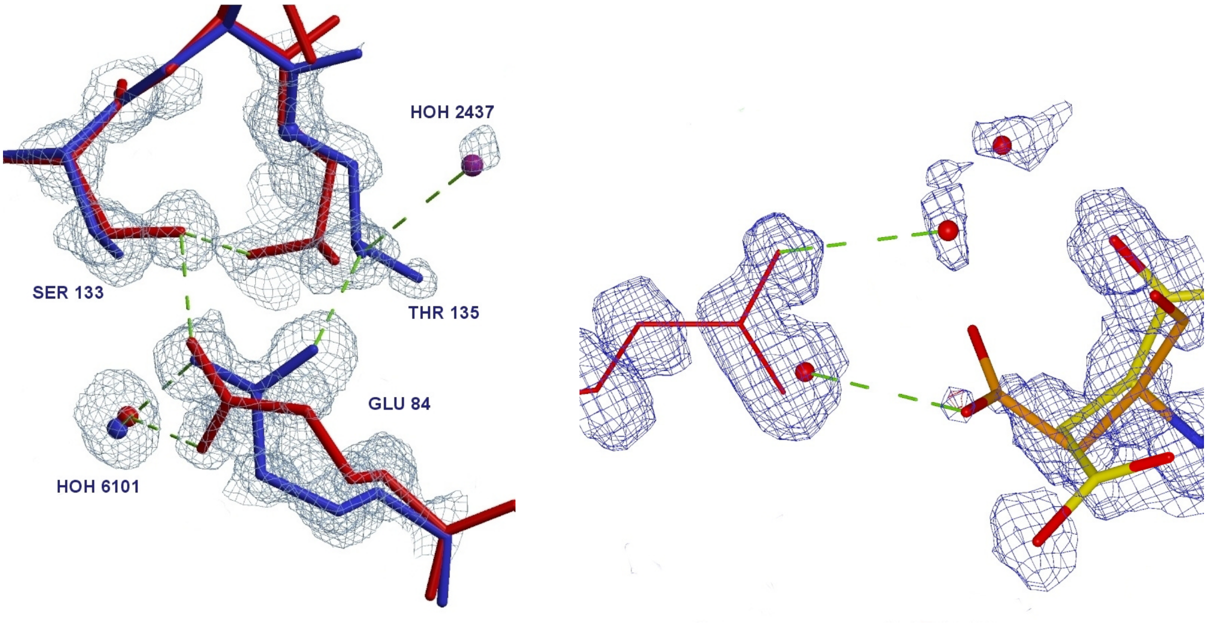
4.1.1.4 Refinement of disordered residues

In the final structure alternative positions were modelled for 23% of the amino acid residues. In general double conformations can be classified as pure side chain disorder, pure backbone disorder and a combination of both. Of the 71 disordered or partly occupied protein residues, 46 belong to the first class, 8 to the second and 13 to the third.

Many of the disordered residues are connected to each other via direct hydrogen bonds or water networks and are therefore correlated. Table B.1 of the appendix lists all free variables refining the occupancies of disordered groups (side chains, whole residues, waters and smaller protein regions of several residues, called ‘clusters’) and the assignment of one or more of these species to each variable.

The presence of two alternative conformations for whole clusters of residues, alternatively connected by hydrogen bonds, is a prominent feature of the hAR2 structure as determined in the scope of this work. Fig. 4.6a shows a part of such a cluster, refined with free variable #2. Five of these larger clusters with at least three protein residues involved were found in the structure, plus several groups of disordered side chains connected to partly occupied waters.

A certain conformer position of one residue may cause the avoidance of a colliding position for other conformers. This type of connective disorder is, unlike in the cluster cases discussed before, not based on attractive forces. In the hAR2 case, it is often observed between side chain types than cannot build hydrogen bonds (other than indirect ones via a water molecule), because they are both of the same type, either both hydrogen donors or acceptors. In the case of Glu71 and Asp134, conformers of both residue side chains, not belonging to the same disorder component, occupy positions that would be otherwise too close for a non-attractive distance. The positions are realized because of hydrogen bonds to partly occupied water molecules (Fig. 4.6b). From another perspective, this state can be described as a repulsion effect between the two negatively charged carboxylic side chains of the same disorder component.



(a) **(b)**

Fig. 4.6 (a): Part of the disordered cluster of residues refined with the free occupancy variable #2. The major conformation (displayed in red) is occupied with 59%. The five subsequent residues 133-137 and the residues 71, 83, 84 plus six water oxygen atoms are participating. **(b):** Residue Glu71 with one and residue Asp134 with both of its side chain conformations displayed. The carboxylic side chains facing each other belong to opposite disorder components.

A third "principle" of disorder, not explainable with alternative intramolecular contacts, is the presence of conformationally similar, but positionally shifted disorder components. This phenomenon, presumably caused by positional entropy in geometrically or chemically less restrained protein regions, is mainly observed for parts of the hAR backbone, but for some residues, in particular Trp295 and Lys296, also the side chains are affected.

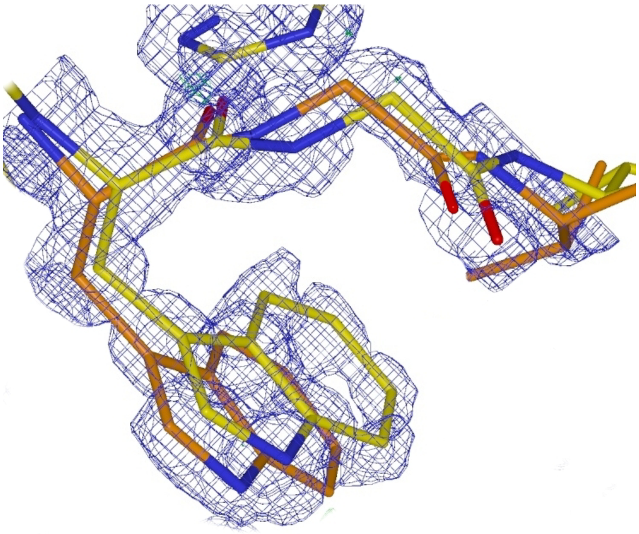


Fig. 4.7 The shifted disorder components of Trp295 and some additional main chain part.(the backbone disorder is continuing up to Ala299).

The disordered hAR2 residues and the correctness of their modeling are dealt with more closely in section 4.1.4, where their verification using the MAD map is discussed.

4.1.1.5 Floppy protein regions

As long as positional entropy can be described by two alternative, discrete states, the modeling of shifted double conformations will allow a sufficiently good refinement which reduces high B-values as well as difference density peaks. In case of hAR2, two especially floppy regions were found, for which modeling of disorder proved to be impossible. These are the C-terminal region, and the slope region of a larger loop on top of the hAR barrel (in fact, in proximity to the C-terminus).

The sequence of hAR2 ends with the amino acid residues His312, Glu313, Glu314 and Phe315. Especially the last three residues did not fit the electron density well. σ_A density is almost completely missing for the four last side chains (Fig. 4.8) and most parts of the protein backbone except for the terminal carboxylate group itself. Moreover, there are significant peaks of negative Fo-Fc difference density at the carboxylate group of Glu313 and at the main chain atoms Cα(314) and N(315). Positive difference electron density peaks, indicating alternative positions of the side chains or a different trace of the backbone, were missing. A look at the experimental electron density map later on confirmed the fact that the whole C-terminal region is positionally very flexible so that a defined electron density shape of a fixed position cannot be observed. The last three residues as well as the side chain of His312 were removed from the model in step 11 of the refinement process and ignored from then on.

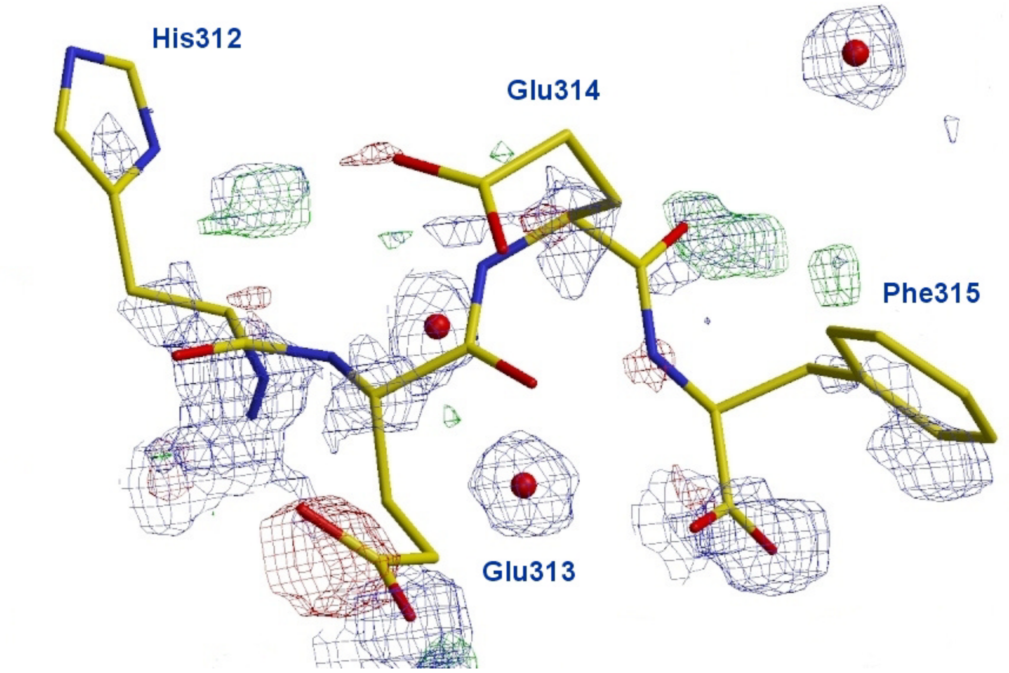


Fig. 4.8 The C-terminal hAR2 region at a structure determination stage before the residues 313-315 were removed from the model. The map contour levels are +/ - 3 σ for the difference density map and 1 sigma for the σ_A map.

The residues Pro222, Glu223, Asp224 and Pro225 form the slope (or hairpin) part of a 20 residue loop (210-229), located on top of the hAR $(\beta\alpha)_8$ barrel (Fig. 4.9a). The slope fold is due to the fact that there are backbone turns at both *cis*-prolines. Loops, lacking secondary structure hydrogen bonds, are in general flexible protein regions. Additionally, the slope part – more precisely residues Lys221 to Asp224 – are exposed to the bulk solvent, being located in a cleft between three symmetry equivalent hAR molecules (Fig. 4.9b).

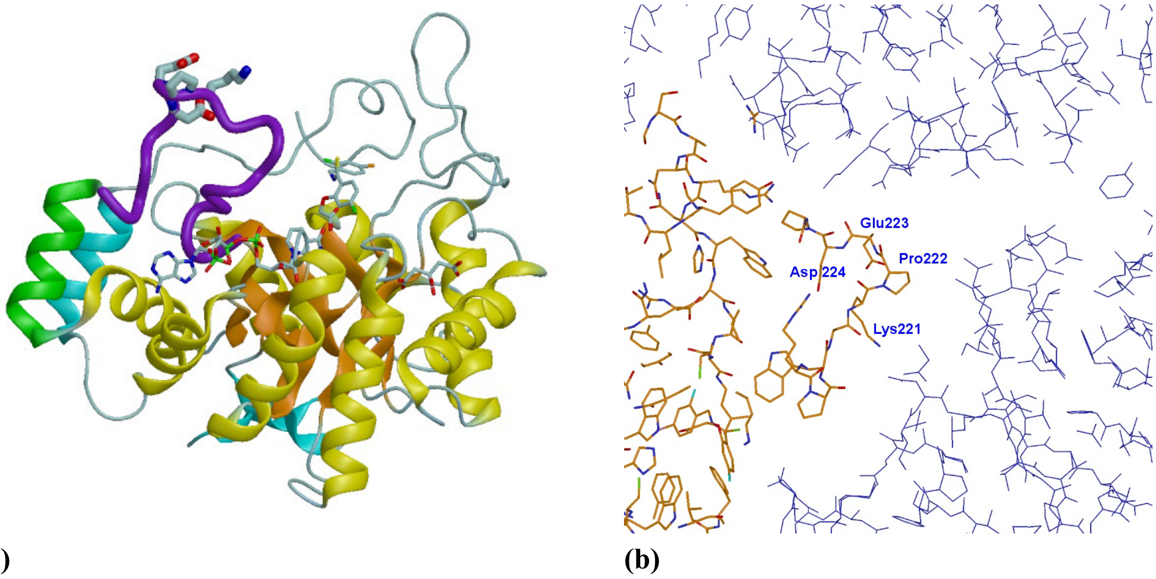
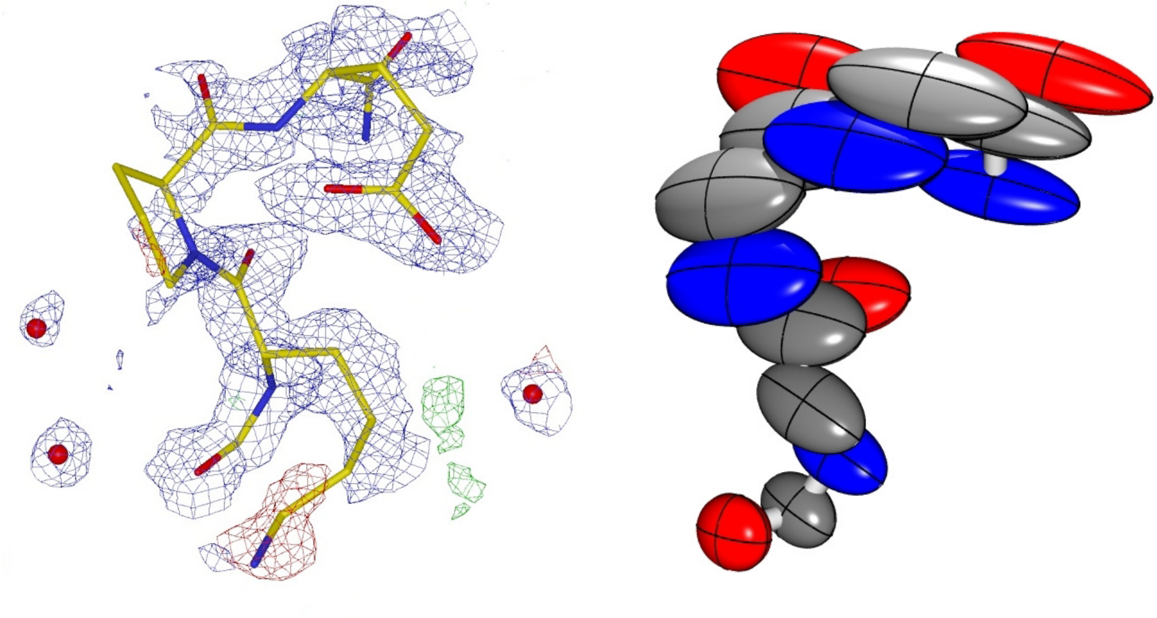


Fig. 4.9 (a): Location of the residue quadruple 221-224 in the hAR2 structure, located on the purple-coloured loop. **(b):** The location of the residues at the interface of three hAR2 molecules.

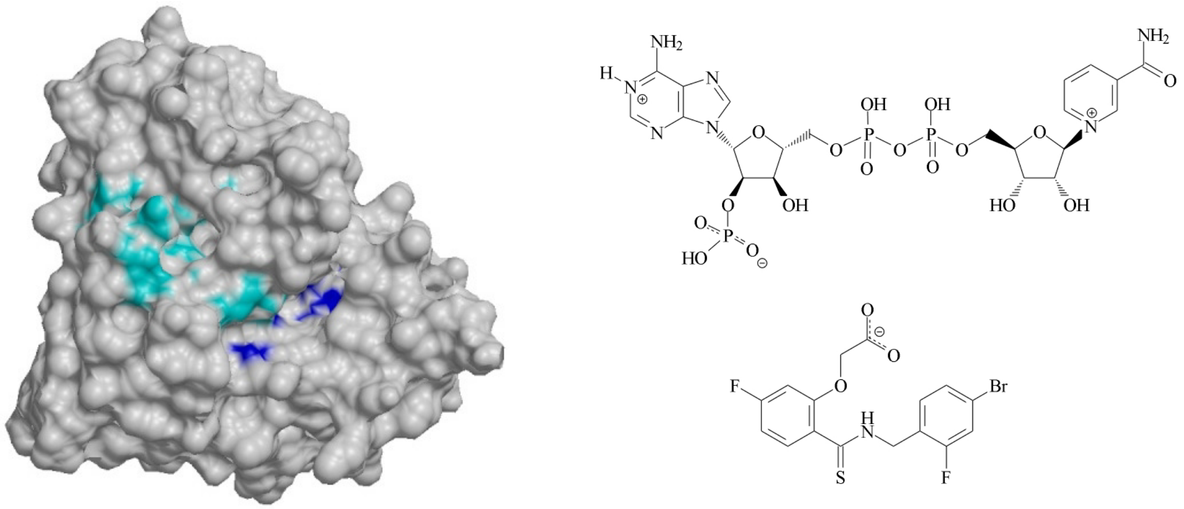
These two circumstances favour the positional entropy of the four residues, as indicated by remarkably high B-values, being 36.2 Å^2 on the average for all side chain atoms and 31.7 Å^2 for all main chain atoms. In contrast to the C-terminus, there is no bigger gap or even complete absence of σ_A electron density, but just a somewhat blurred shape with some smaller density gaps (Fig. 4.10a). Therefore, usual multiple-conformation disorder seems less probable than a multiple shift of positions. The four-residue slope part does not need to be removed, yet it seems hardly possible to model twofold disorder, either.



(a) **(b)**

Fig. 4.10 The residues Lys221, Pro222 and Glu223. **(a):** Model with σ_A map at one sigma contour level (blue) and F_o-F_c map at +/- 3 σ level (green, red). **(b):** Thermal ellipsoids representing the main chain atom ADPs, same perspective as left.

4.1.1.6 The hAR2 active site



(a) **(b)**

Fig.4.11 (a) The surface of hAR2 with residues interacting with the coenzyme NADP$^+$ highlighted in cyan and residues interacting with the ligand highlighted in blue. **(b)** Chemical fomula of NADP+ (top) and of the ligand IDD594 (bottom).

As explained in the structural overview, coenzyme and ligand are neighbouring each other at the top of the hAR $(\beta\alpha)_8$ barrel. Fig. 4.11 shows the clefts on the surface of the apo model at the interaction sites of the two small molecules. From the structure-biological point of view, the question about the protein fold in the contact region is important. For the crystallographer, the question about consequences emerging for the protein residue geometry, especially with respect to backbone torsion angles and possible restraint violations is an interesting one as well.

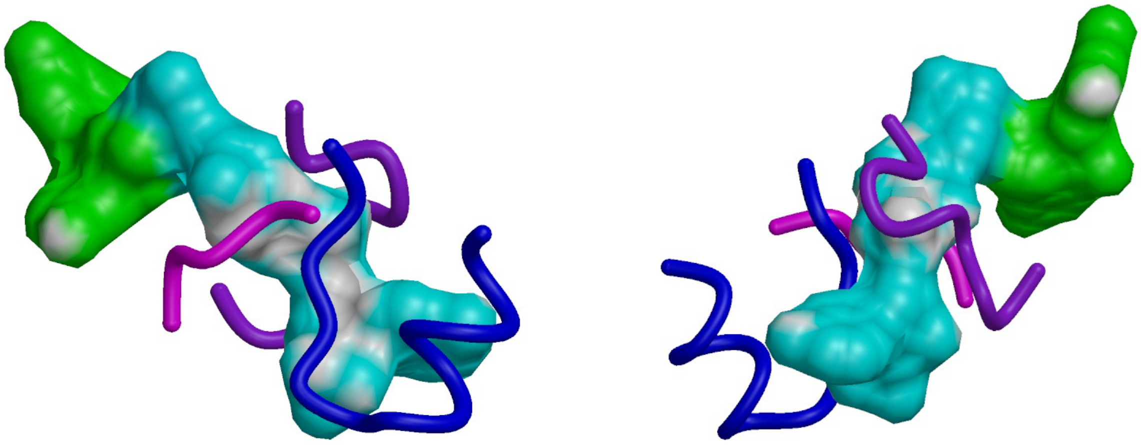
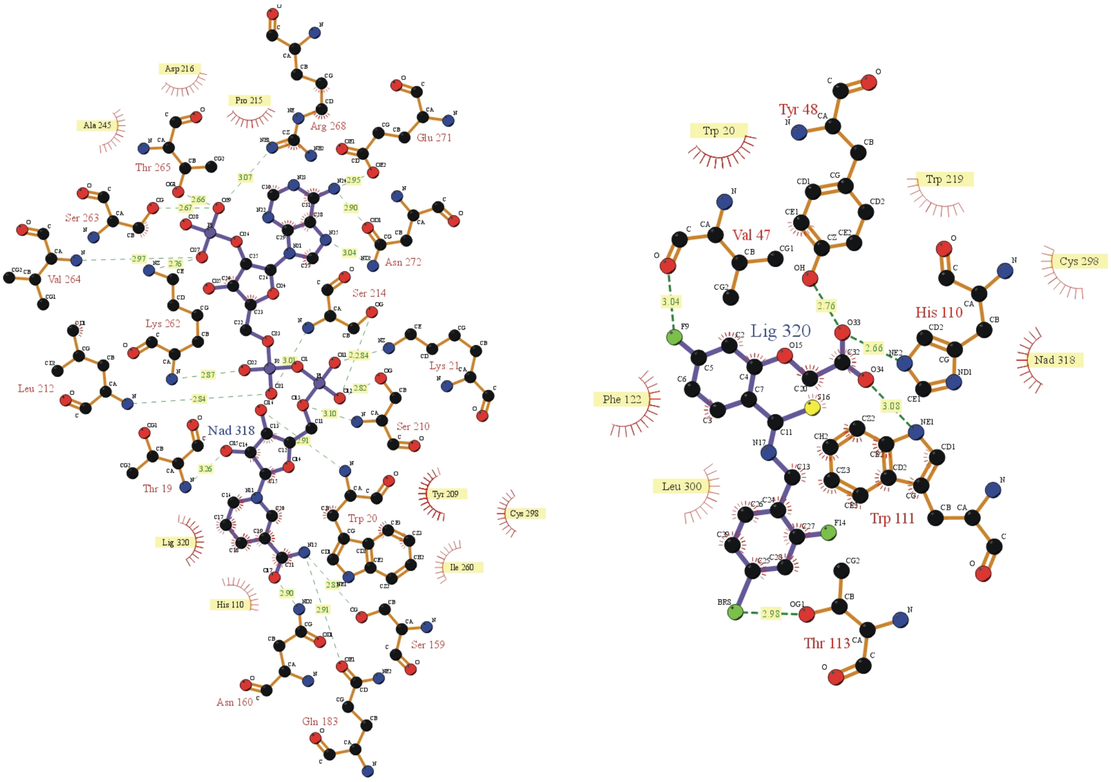


Fig. **4.12**: The outer shape of the ligand and coenzyme molecules and their environment of close parts of the protein backbone fold. The molecules are displayed in surface mode, $NADP_+$ being coloured cyan and IDD594 coloured green. Fragments of the hAR2 main chain containing the interacting residues (as presented in Fig. 4.2) are displayed as tubes of different colours. Blue: residues 260 – 273.. Purple: residues 209 – 218. Magenta: residues 17 – 23. Both images show the same arrangement from different perspectives.

Both the ligand and the coenzyme molecule are surrounded by protein loops (Fig. 4.12). Characteristic turns of the backbone can be observed. For example, the sequence of residues 260-273 is turning twice in order to flank the adenine side of the coenzyme. The turns required for this are established by the residues Pro261 and Pro266. Consequently, these two residues exhibit non-standard geometries, in particular a peptide bond deviating from planarity, as indicated by restraint violations(see Table A.2). The same is true for Pro215 and Pro218, belonging to the loop at the opposite site of NADP$^+$.



(a) **(b)**

Fig. 4.13 Schematic representation of all hydrogen bond interactions between **(a)** hAR2 – NADP+ and **(b)** hAR2 – ligand, given with distances.

The coenzyme is directly interacting with 16 hAR2 residues. Only the nicotineamide moiety, being the hydride donor for the substrate reduction (see introduction), is partly accessible (Fig. 4.13a). The nicotinamide ring is positionally fixed by the hydrogen bonds Asn160_Nd – O17 (2.90 Å), Ser159_Og – N12 (2.85 Å) and gln183_Oe – N12 (2.91 Å).

The inhibitor molecule is less buried (Fig 4.13b). There are four direct intermolecular contacts, the hydrogen bonds established by the carboxylic inhibitor “head”, Tyr48_Oη–O33 (2.76 Å), His110_Nε-O33 (2.66 Å), Trp111_Nε–O34 (3.08 Å), and the bromine-oxygen contact to Thr113 (2.98 Å). Additionally, a π-stacking between the bromo-substituted phenyl ring of IDD594 and Trp111 is found (Fig. 4.14). The inhibitor also establishes an intramolecular hydrogen bond, O34–N17 (3.04 Å) responsible for (or resulting from) the special conformation of the molecule, fitting into the protein binding pocket.

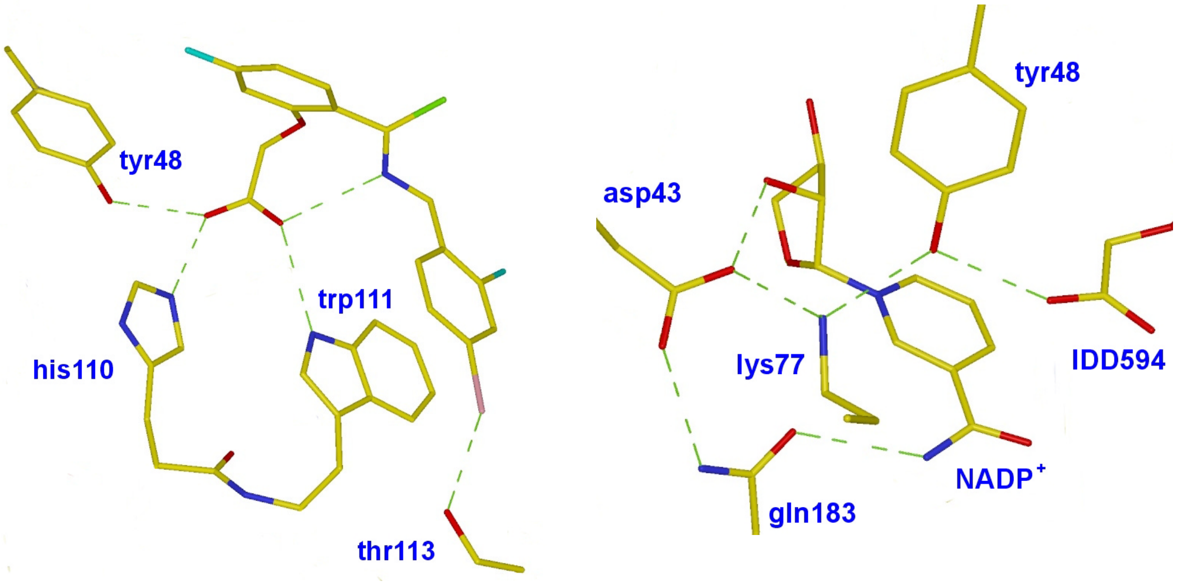


Fig. 4.14: **(a)** Conformation and interactions of the inhibitor molecule, **(b)** additional interactions between residues involved in the catalytic mechanism of hAR2. Lys77 is the proton "substitutor" for Tyr48 according to the proposed reaction mechanism (Fig. 1.3)

The carboxylic moiety of IDD594 is occupying the position of the aldehyde functionality of substrates, close to the nicotinamide system (Fig. 4.14b), but due to the chemical properties of the inhibitor, no reduction occurs and the binding is irreversible. This explains why IDD594 is a non-competitive inhibitor.

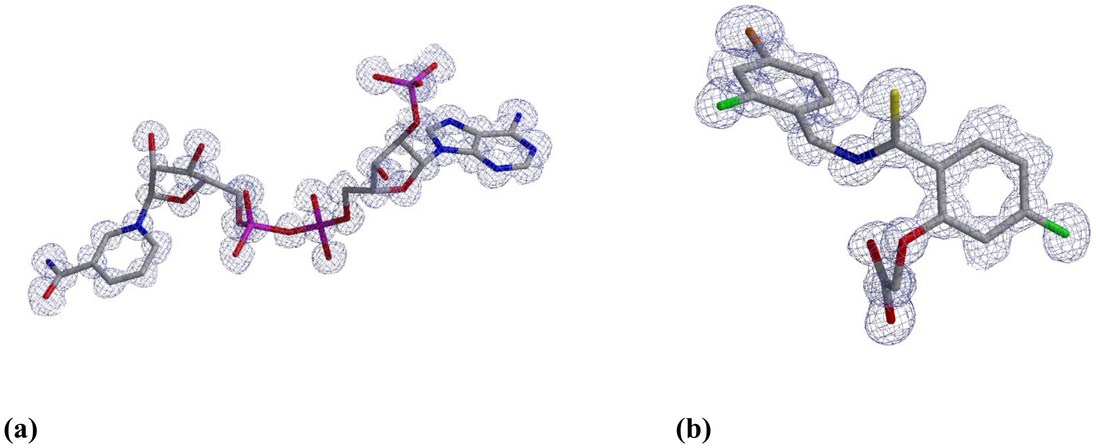

(a) (b)

Fig. 4.15 : σ_A density at a contour level of 1 σ, as observed at **(a)** the NADP molecule and **(b)** the IDD594 molecule.

The active site is a positionally very fixed region, and the electron density covering the two molecules is very defined (Fig. 4.15). The ligand as well as the coenzyme and the interacting amino acids have B-values below the protein average (Table 4.1). The mean B-value for all active site atoms including the participating residues is 5.2 Å^2. Thus, it is not surprising that multiple conformations are almost completely missing at the active site.

One exception is residue Cys298. Due to changes in activity and ligand binding upon mutagenesis (C298S), this residue is believed to be implicated in the catalysis as a regulatory group (Petrash et *al.* 1992). The residue is located at the “free” side of the nicotinamide moiety, where the hydride transfer normally takes place. In the present structure, Cys298 was found to be disordered.

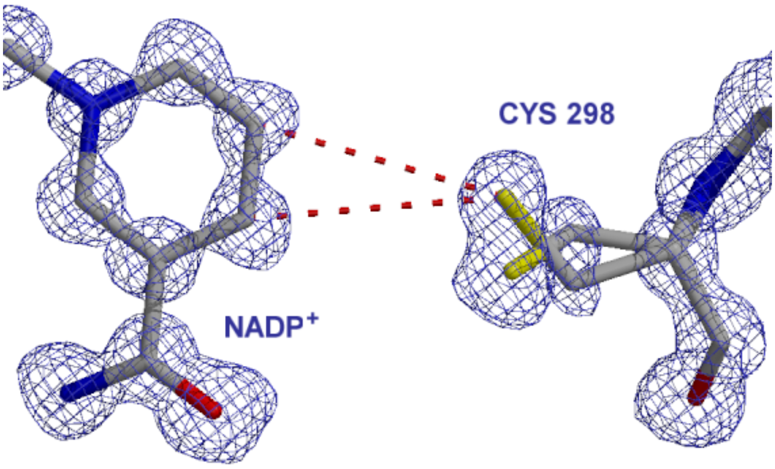


Fig. 4.16: Residue Cys298 interacting with two phenyl ring carbon atoms of the nicotinamide side of the coenzyme.

Interestingly, only the minor conformation of the Cys298 side chain (occupancy 43%) shows contacts (3.76 Å) with the coenzyme molecule, thus sterically blocking the transferable hydrogen atoms (Fig. 4.16). It seems reasonable that, assuming a regulating function of Cys298, the side chain acts as a switch allowing the reduction reaction to take place in its major conformer position and preventing it with its minor conformation.

4.1.1.7 Water molecules in the hAR structure

The total (theoretical) solvent water content of hAR is 43%. This corresponds to a water oxygen atom number of about 2200. While the low solvent content is another explanation for the good diffraction of hAR crystals, it has to be kept in mind that an important fraction of solvent water is completely disordered, *i.e.* no atomicity of the electron density distribution is found in this so-called bulk solvent region of the crystal, which is the space between the roughly spherical shaped protein macromolecules. The closer water molecules are located towards the protein, the more likely become defined water networks connected by hydrogen bonds. There is of course a gradual transition from bulk solvent water to well ordered, regularly coordinated and thus positionally fixed water molecules. This fact is reflected by the wide range of water B-values or by the varying reduced occupancies of water oxygen atoms, respectively. Highly resolved electron density maps allow the modeling of partly occupied waters. In case of hAR, the occupancy sum of water oxygen atoms corresponds to a number of 546 fully occupied atoms, which is about a quarter of the total solvent fraction. In addition to the rather low solvent fraction itself, the strong crystal diffraction is even more favoured by this relatively high quota of more or less ordered waters.

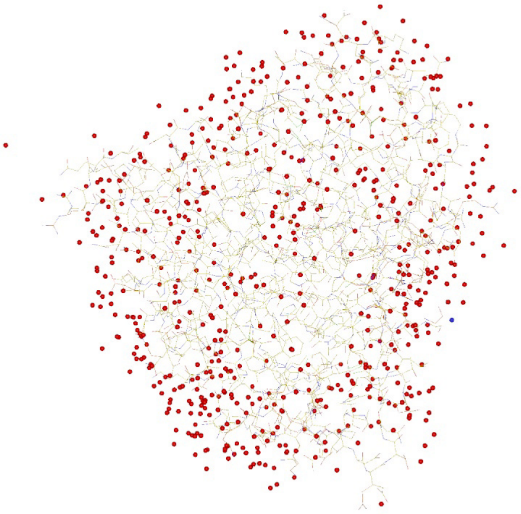


Fig. 4.17: The ensemble of water molecules modelled in the hAR structure

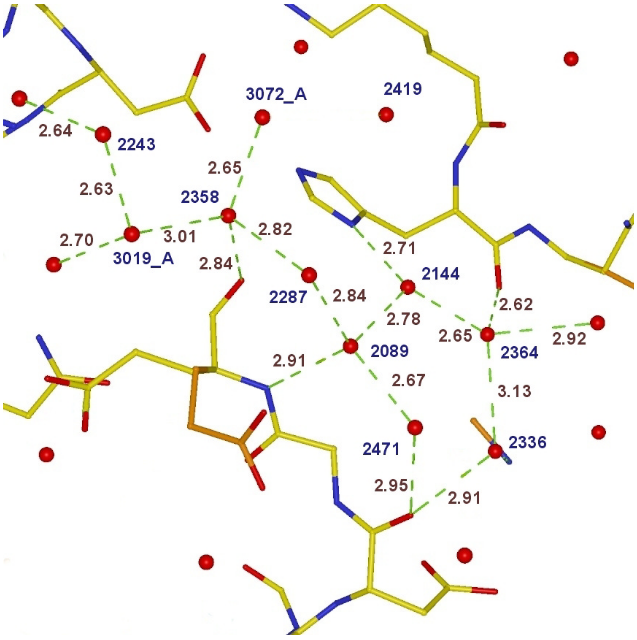


Fig. 4.18: Part of a network of fully occupied water molecules.

As described in the methods section, virtually all fully occupied water molecules were modelled in the first refinement steps. They are building networks which are mainly located at the surface of the protein, between the side chains of polar amino acid residues.

Parts of the networks are overlapping, leading to water positions too close to belong to the same network. These water molecules were mostly found in the late structure determination steps, after inspecting new difference electron density peaks or after the revision of previously fully occupied water molecules, some of which exhibited negative difference density and (or) very high B-values of more than 50 A^3.

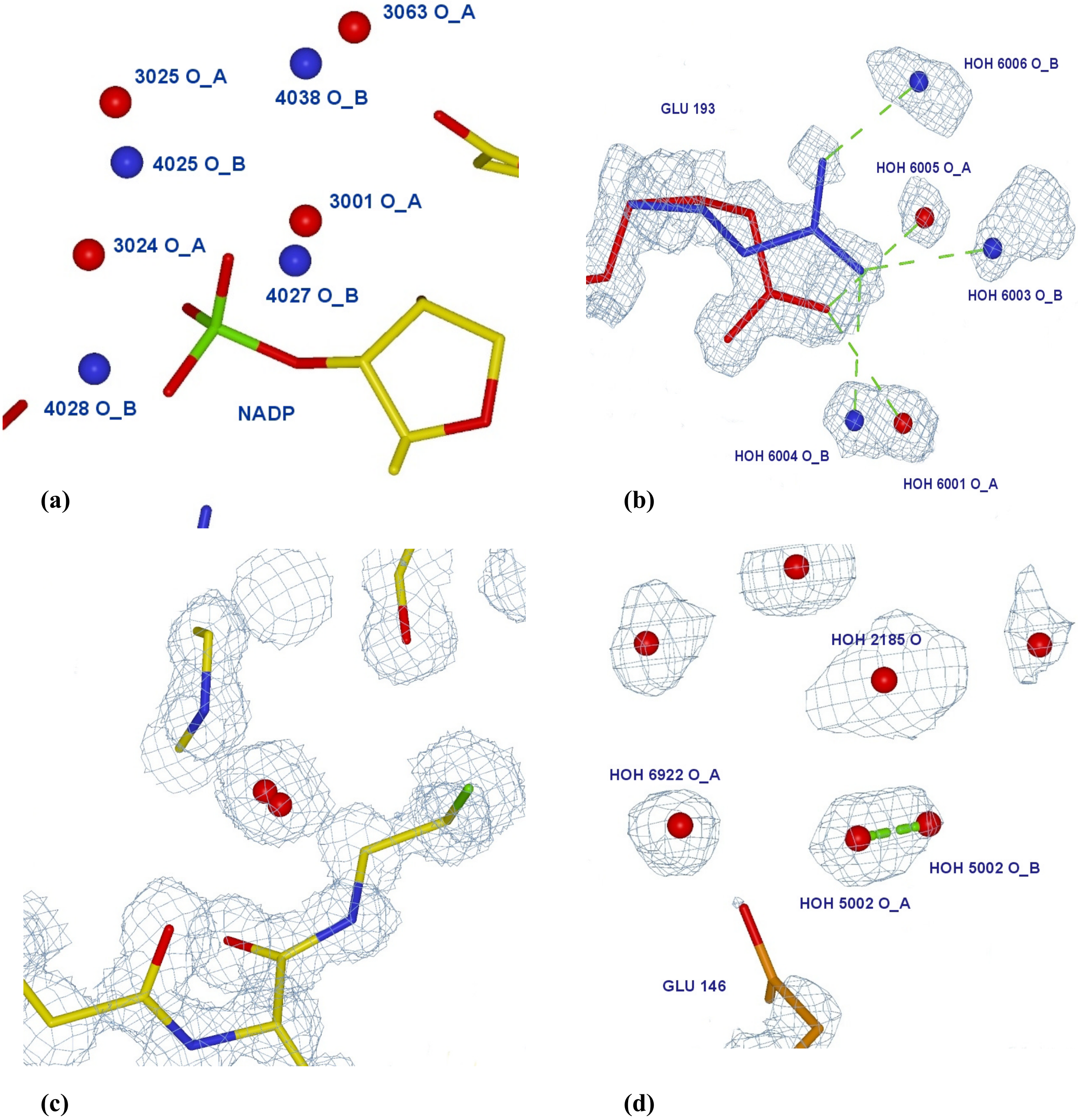


Fig. 4.19: Types of partly occupied water oxygen atoms **(a)** A double network (close the mono-phosphate group of the coenzyme) refined with alternative parts, each with 50% occupancy. **(b)** Water oxygen atoms connected to the two conformations of residue Glu193 **(c, d)** Separate pairs of water oxygen atoms with individual occupancies, summing up to unity. They exhibit a common ellipoid-shaped σ_A electron density. As example **(c)** shows, some of the pair atoms have almost the same location after refinement.

Water oxygen atoms of the second type were given fixed occupancies of 50%. Separate pairs of very close water oxygen atoms represent the third type of solvent water, often found to have a common ellipsoid-shaped Sigma-A density peak. They were given the same residue numbers and an exclusive free occupancy variable. The last type of water molecules are those connected to protein side chain disorder. They build hydrogen bonds to either of the two side chain conformations, leading to what can be called a small cluster (or they are part of a larger cluster consisting of more than one protein residue). In any case, they were given the same free occupancy variable as the respective residue(s).

type	description	atoms found	residue numbers	occupancy	min. B [Å^2]	max. B[Å^2]	mean B[Å^2]
I	normal network	426	2001 –	1.0	3.84	50.67	21.65
II	double-network	114	3001 – (conf. A)	0.5	4.70	44.73	20.16
		47	4001 – (conf. B)	0.5	6.16	33.73	21.13
III	disordered pairs	14	5001 – (A, B)	variable	3.07	18.69	11.41
IV	side chain-connected	64	6001 – (A, B)	variable	5.66	38.53	18.45

Table 4.4: Overview of water modeling and refinement.

Table 4.4 gives a summary of water molecule types and the number of water oxygen atoms found for each type. A total number of 665 water oxygen atom positions have been found. Most of the atoms (64%) belong to the not disordered main network. The mean B-values are comparable for all water types with fixed oxygen atom occupancy. A value of about 20 Å^2 is acceptable, and only one water exceeds the formal threshold of 50 Å^2. The B-values are significantly lower for those atoms with occupancies refined variably, in particular for the disordered water pairs of type III. This is not surprising, as a variable refinement is in general more precise. Theoretically, the same method could be applied to the so-called double water network, but the decision whether to refine those waters as one big group or in several subgroups (like the 2 x 4 atoms in Fig 4.19a) is not a trivial one. Too big clusters of atoms belonging to one group with a common variable prevent a more differentiated occupancy refinement, but the assignment of water oxygen atoms to seperated groups is in most cases not as obvious as here.

4.1.1.8 Visibility of hydrogens in the difference density map

Before activating the hydrogen atom placement and refinement in SHELXL (see methods part), the atomic positions were checked, using the respective difference electron density maps in XFIT. As shown exemplarily in Fig. 4.20, the results are rather inconsistent even for the same type of hydrogen atom in the same region of the protein. In general, hydrogen atoms connected to Cα atoms are best visible, but even though, there are difference density peaks missing or very small for Cα hydrogen atoms in the β-sheet part of the structure shown in Fig. 4.20b (the residues 182 to 184 and 206 to 208 are displayed). The size of the peak shapes differs corresponding to the peak heights. Therefore, lacking consistency, the recognition even of Cα–hydrogen atoms is not completely possible at the usual level of 3 σ for Fo-Fc density, and in some cases neither at lower contours. It was also be found that nitrogen-bound hydrogen atoms have a lower mean difference density peak level than the Cα ones, and for side chain hydrogen atoms the level is even lower. Fig. 4.20a emphasizes that 1.2 Å map resolution is less sufficient than 0.9 Å resolution for hydrogen atom visibility – exactly like one would expect for hydrogen atoms, having a covalent bond distance of about one Å in crystal structures. However, the generally decreasing data quality at the very high resolution border (see chapter 3.1.3) had given rise to the question whether slight resolution truncation could improve high-resolution dependent map details. At least for the hydrogen peaks in not experimentally phased F_o-F_c difference density maps the answer is that very highly resolved structure factors are essential despite their reduced accuracy.

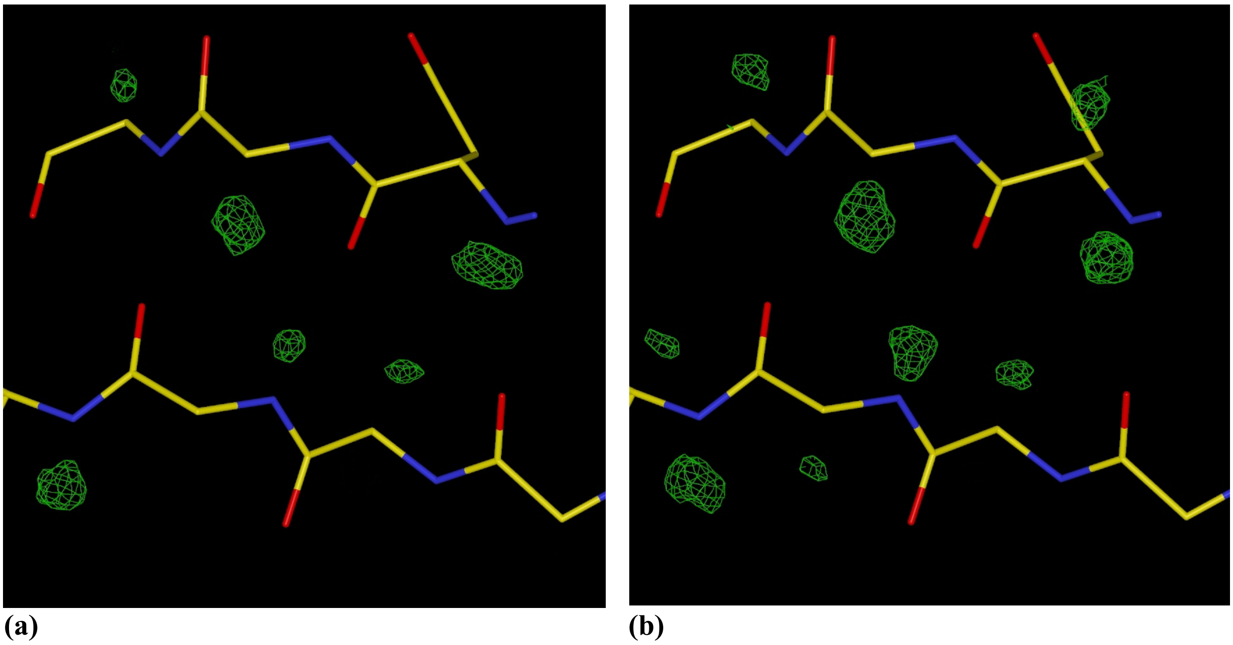

(a) **(b)**

Fig. 4.20: The appearance of backbone hydrogen atom omit positions in Fo-Fc difference electron density maps contoured at 2.5 sigma. (a) at 1.2 Å. (b) at 0.9 Å resolution for an examplary, well ordered β-sheet part of the model.

Polar hydrogen atoms connected to the oxygen atoms of tyrosine, serine, threonine, and the phosphate groups of NADP$^+$ as well as all possible four hydrogen atoms of the five-membered imidazole ring of histidine were not modelled. Whereas the check for main chain hydrogen atoms has more formal reasons, the visibility of polar side chain hydrogen atoms can be useful to characterize hydrogen bonds possibly being important interactions in the active site region. The reduction mechanism of the reductase class enzymes, based on the coenzyme NADP$^+$ as hydride source, is based on proton transfer. Thus, the presence or absence of hydrogen atoms at key positions of the coenzyme and the protein residues directly bound to NADP$^+$ would help to reveal the exact mechanism of the substrate reduction process (see introduction). In general, the scattering contribution of hydrogen atoms beyond 1.2 Å is not significant – however the ADPs of the bound heavier atoms (C, N, O) are determined more precisely at higher resolution, preventing them from covering the hydrogen electron density.

Unfortunately, the visibility of polar hydrogens atoms in general is even poorer than that of non-polar ones. The main reason might well be the more delocalized state of the single hydrogen electron, which is tendencially drawn even further to the donor atom along the weakened (and prolongated) covalent bond. Summing up, it can be stated that the overall quality and resolution of hAR2 structure factors, although very high, is not sufficient to allow completely satisfying hydrogen atom identification.

Further results and a more detailed listing of the presence or absence of difference electron density peaks related to hydrogen atoms are presented and explained both in the previous chapter about the active site interactions and in the following chapter, where hydrogen atoms in experimentally phased maps are discussed.

4.1.1.9 Special features of the hAR2-IDD594 structure

Every heavy atom position of the hAR2 model (i.e. the positions of the six methionine selenium atoms and the inhibitor bromine atom) exhibited large negative peaks in the F_o-F_c type difference density maps during the earlier refinement stages. The first idea was to look for a common reason for this phenomenon. Such an explanation would account for the fact that all affected atoms have about the same number of electrons and an electron density considerably higher than the rest of the structure. Following this argumentation, the up-weighting of protein electron density by the SHELXL bulk solvent model would affect heavy atoms in a way that exceeds experimental evidence. In other words, the calculated structure factors would become artificially over-weighted relative to the observed ones, and therefore, difference-density would become significantly negative. To check this theory, comparative SHELXL refinements with and without the *SWAT* bulk-solvent correction command were carried out. Additionally, the same two refinements were done against low-resolution-truncated data (ranging from 6 to 0.9 Å only). However, the negative difference peak values did not change significantly in any of the cases.

An alternative explanation for negative difference density would be radiation damage. Assuming the dissociation of weak carbon-bromine and carbon-selenium bonds and the subsequent elimination of bromide and selenide species, the true atom occupancies at the relevant positions would be lower than actually modelled (*i.e.* less than fully occupied). As the development of usual X-ray damage can be regarded as a continuous process, for which a gradual (often linear) decomposition of the weak bonds is observed, the phenomenon of negative difference density should be more accentuated towards the end of a diffraction experiment. In case of the data collection on the hAR-IDD594 Se-Met derivative crystal, like for most MAD experiments, peak data were collected first and high-energy remote data last – but different regions of the crystal were exposed for each scan. Therefore, by comparison of the refinements against all 1.5 Å data subsets, no significant differences were found with respect to the difference electron density peaks (Table 4.5). To prove radiation damage (within the single scans), zero-dose extrapolation should be applied to the data.

data subset	Fo-Fc minimum [sigma units] for 100% bromine occupancy	refined bromine occupancy
Peak	-10.4	73.4%
inflection point	-7.6	75.8%
high-energy remote	-12.3	73.3%

Table 4.5: The negative density phenomenon at the bromine position for different data source F_o-F_c maps

Apart from the first two explanations, individual reasons for the negative difference density had to be checked. At the selenium atom positions, the presence of two atom types – a majority of selenium atoms, and a minority of sulphur atoms, explainable by incomplete Se-Met derivatization, was proposed. The refinement of atom type disorder (which is of course connected to a slight positional disorder), using free occupancy variables, lead to the following results:

Residue	selenium occupany	sulfur occupancy
Met12	89.6%	10.4%
Met144	95.4%	4.6%
Met253	87.2%	12.8%
Met285	92.8%	7.2%

Table 4.6: The refined occupancy compositions of selenium-sulfur disorder in the four positionally fixed methionine residues. For the multi-conformational residues Met1001 (N-terminus) and Met168, atom type disorder was not modelled.

Non-positive definite ADPs for the sulfur atoms were observed initially, but could be prevented later by the application of the SHELXL *EADP* instruction, making the sulphur ADP values equal to the selenium values.

In case of the IDD594 inhibitor molecule, the incomplete incorporation of the bromine functionality is chemically very unlikely. Additionally, mass spectra have proven the existence of bromine in the molecule to be 100 per cent. Howard et *al*. (paper submitted to proteins) have proposed a polarization effect in the bromine environment. Looking at the close bromine-oxygen contact of 3.0 Å between IDD594 and Thr113, a electron density transfer towards the oxygen atom, having the higher electronegativity, seems to be possible.

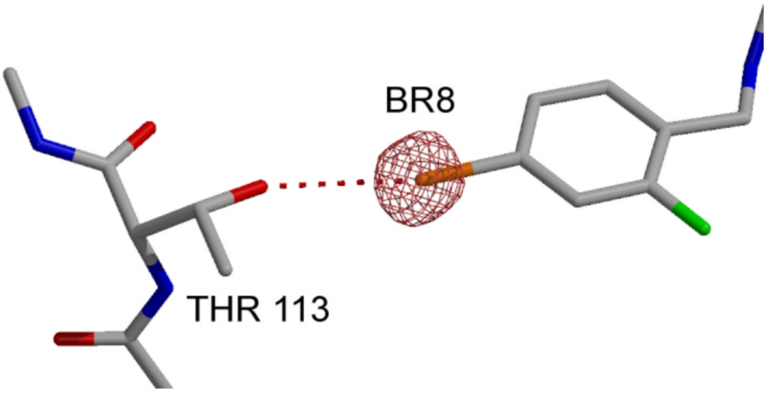


Fig. 4.21: The negative F_o-F_c electron density peak at position Br8 of the inhibitor molecule and the interaction to the oxygen atom of Thr113.

A negatively charged threonine Oγ atom and a positively charged bromine atom would result. However, this simple assumption would require the presence of positive difference electron density at the oxygen atom as proof. Furthermore, the refined bromine occupancy of 73.3% would correspond to a lack of nine bromine electrons. It seems unlikely that this large amount of negative charge can be located at one acceptor atom alone.

4.1.2 Evaluation of the experimental phases derived from hAR data

The features of the anomalous hAR datasets with respect to signal-to-noise ratio and inter-wavelength correlation have been presented in the methods chapter. The very good overall values as well as some relatively bad (but still acceptable) regions at 3.8-3.2 Å and beyond 1.1 Å are to be emphasized.

All graphs shown in this chapter are based on a determination of differences between experimental SHARP phases and calculated phases from the refinement (step 10), as explained in chapter 3.1.7 Unless explicitely defined otherwise, phase errors discussed in this and the following chapters refer to the F^*FOM-weighted values.

Systematic data errors in the region of about 3.5 Å reduce the anomalous signal-to-noise ratio for all data subsets used and - as it is a wavelength-independent effect - also the correlation between the subsets (see Fig. 3.1). This has consequences also for the experimental phases derived from that data: Indeed, the phase quality for the 3 to 4 Å resolution bin is worse than for lower and higher resolution, as observable in figures 4.22 to 4.26 for the phase error, FOM and map-CC curves.

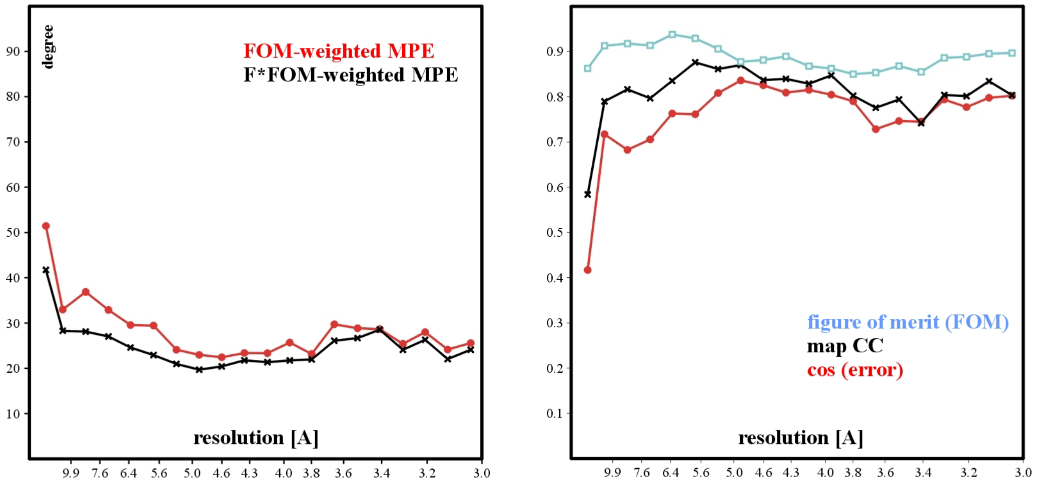


Fig. 4.22: SHARP phase errors (left) and correlation coefficients (right) plotted against resolution [Å] for the heavy atom refinement and protein phase calculation (called "job") using MAD data to 3.0 Å. Correlation coefficients CC are calculated according to Lunin & Woolfson (1993). MPE = mean phase error.

Phase errors are largest for very low resolution data and rise again at 3.2 – 3.7 Å (Fig. 4.22). The curves in the correlation coefficient diagram follow this trend. Apparently due to weaker structure factor amplitudes the two phase error curves as well as the CC and cosine-of-error curves are closer in the 3 – 4 Å region. For the SHARP "job" (heavy atom refinement and protein phase calculation) using limited 3.0 Å data, a overall mean phase error of 23.9° and a CC of 81.4% is obtained.

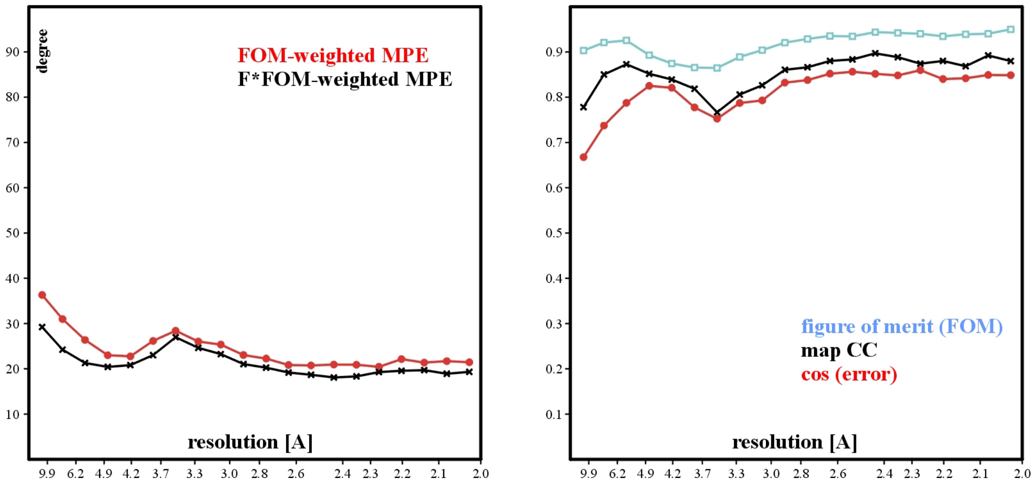


Fig. 4.23: SHARP phase errors (left) and correlation coefficients (right) plotted against resolution [Å] for the heavy atom refinement and protein phase calculation using MAD data to 2.0 A.

The diagrams for the 2.0 Å job (Fig. 4.23) do not significantly differ from the previous results. Because of the different resolution bin scale, the local maxima of the phase errors, as well as the minima of correlation curves at around 3.5 Å both become single peaks. For the SHARP job using data truncated at 2.0 Å, one obtaines an overall mean phase error of 20.9° and a CC of 84.8%.

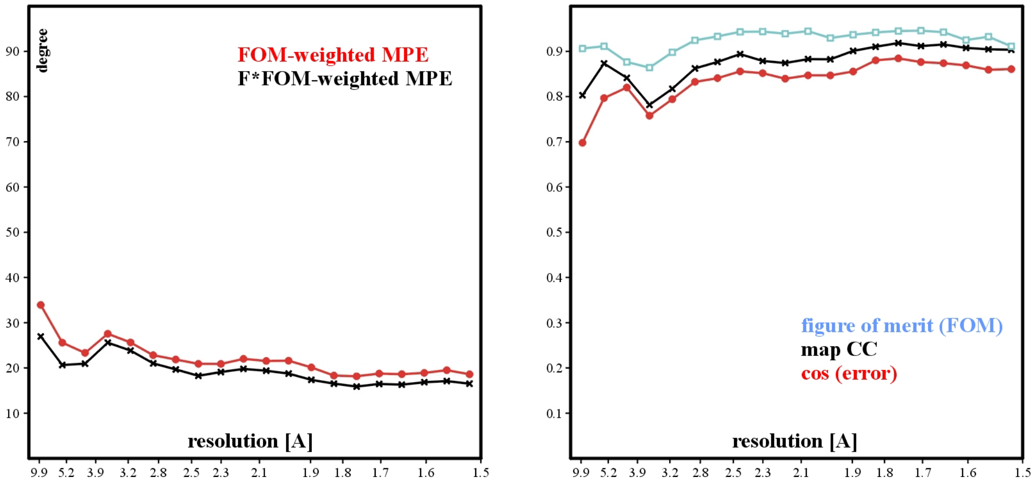


Fig. 4.24: SHARP phase errors (left) and correlation coefficients (right) plotted against resolution [Å] for the heavy atom refinement and protein phase calculation using MAD data to 1.5 Å.

For the 1.5 Å job, the phases in general become more similar to the calculated reference phases (Fig. 4.24). This is reasonable, because the data-to-parameter ratio for the heavy atom model, on which the further calculation of experimental protein phases is based, increases and facilites a more accurate heavy-atom refinement. The results are 19.3° for the overall mean phase error and 86.2% for the CC.

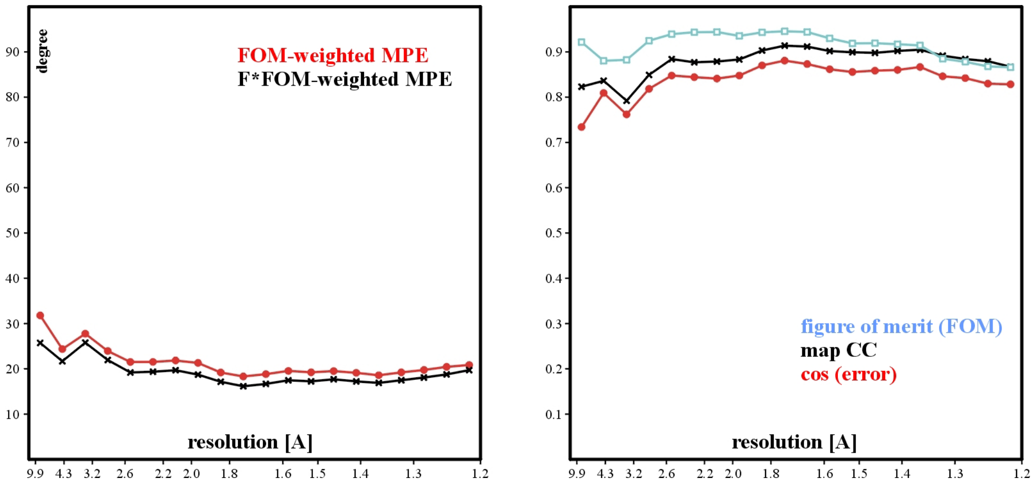


Fig. 4.25: SHARP phase errors (left) and correlation coefficients (right) plotted against resolution [Å] for the heavy atom refinement and protein phase calculation using MAD data to 1.2 Å.

The SHARP refinement job against 1.2 Å data (Fig. 4.25) results again in a slightly better overall phase quality than the previous jobs. The main improvement for the description of the heavy-atom model and the resulting phases is the introduction of anisotropic displacement parameters. The necessity for refinement of the selenium site ADPs has been emphasized previously, in Fig. 3.4. A mean phase error of 19.2° and a CC of 86.2% is obtained. The increase of the phase error curves and the decrease of the correlation curves is only very slow in the data region beyond 1.4 Å.

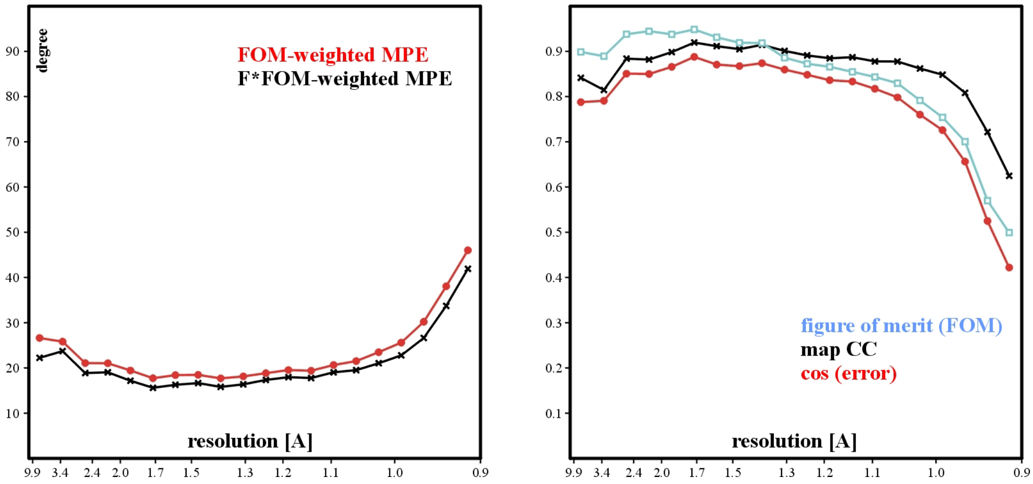


Fig. 4.26: SHARP phase errors (left) and correlation coefficients (right) plotted against resolution [Å] for the heavy atom refinement and protein phase calculation using MAD data to 0.9 Å.

The decrease of anomalous data quality, in particular for the anomalous inter-wavelength correlation coefficients, beyond 1.1 Å (due to higher experimental noise, Fig. 3.1) finally has a direct effect on the experimental phases of the last sharp job. They become worse quite rapidly towards the high resolution limit (Fig. 4.26). This effect, caused by large systematic errors for anomalous structure factor amplitudes, can not be compensated by a maximal data-to-parameter ratio for the heavy-atom model refinement, especially because the dominating heavy-atom contribution to anomalous scattering is affected most by these errors. This justifies the assumption that the accuracy of the selenium sites may suffer from including a large number of *relatively* bad anomalous data into the refinement.

The overall experimental phases are therefore slighly worse than for the previous SHARP job, but still of very good quality. This is indicated by a final mean phase error of 19.7°, and a mean map correlation coefficient of 86.8% (this is the best of all CC values). The curves for the map CC and the cosine of phase error are sufficiently close and have the same trend along the resolution range.

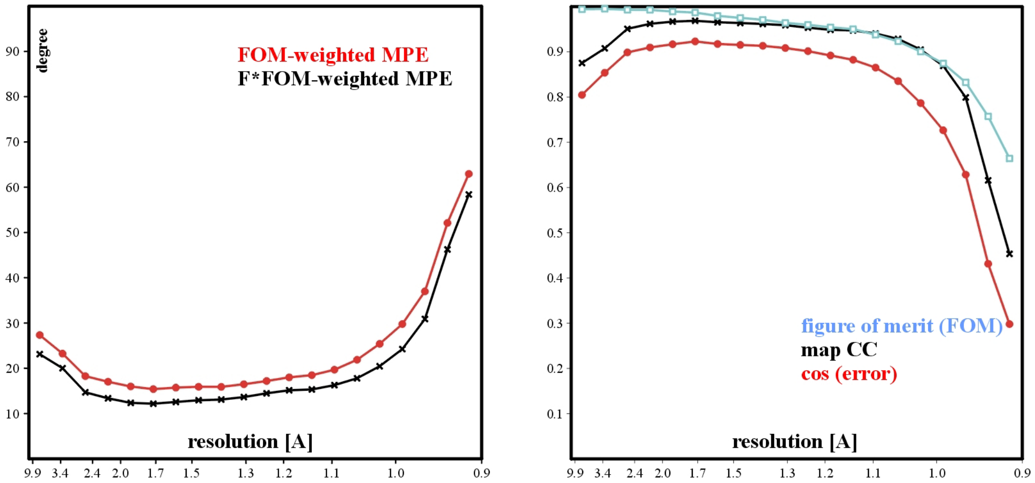


Fig. 4.27: Phase errors (left) and correlation coefficients (right) after the concluding density modification procedure with Solomon (based on the previous SHARP phases for MAD data to 0.9 Å.)

The effect of density modification with SOLOMON on the agreement of the experimental phases to the ones calculated from the model can be described by a general improvement of phase error and correlation coefficient (Fig. 4.27). The results obtained are the best values on the whole, being 17.4° for the phase error and 92.9% for the CC, respectively. The distribution of these values within the resolution bin is however different from all pure SHARP output. Interestingly, the phases on the very low resolution edge (> 6 Å) deviate slightly more from refinement phases than before. Especially for these phases, where the contribution of bulk solvent is largest, a quality improvement by density modification should be expected, if solvent flattening is applied. On the other hand, this does not necessarily have to cause a closer agreement to phases resulting from model refinement. It may well be that the assumptions for the SHELXL bulk solvent model and the SOLOMON solvent-flattening algorithm differ, which results in a larger phase error. At the high-resolution edge, the CC and cos(error) values are worse than after SHARP (it has to be kept in mind that SOLOMON is actually not intended for a resolution higher than 1.5 Å), but the FOM value is better. The FOM values determined by SOLOMON are very close to unity up to 1.2 Å, which seems to be an over-estimation.

4.1.3 Analysis of the experimental electron density maps

Map comparison can be done qualitatively by visual inspection of maps or quantitatively by calculating correlation coefficients serving as real-space similarity indices. Both methods were applied for these studies.

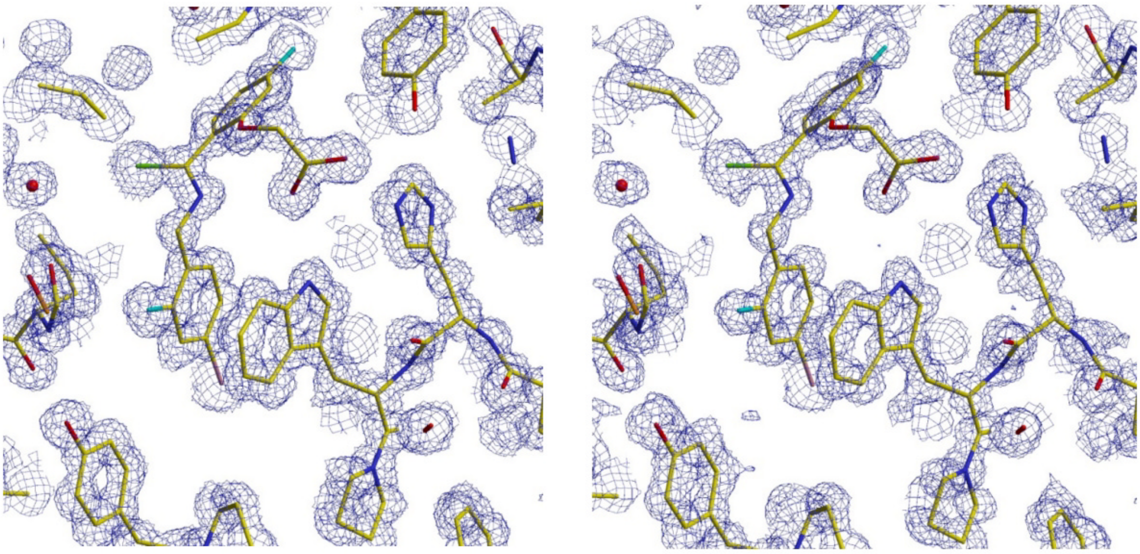

(a) (b)

Fig. 4.28 **(a)** σ_A map and **(b)** experimental FOM-weighted F_o / phi_{MAD} map, both contoured at 1 σ. The exemplary part of the hAR model consists of the inhibitor molecule and the surrounding protein loop region.

Real space map comparison with MAPMAN lead to a calculated overall electron density correlation coefficient of 83%, underlining the good visual similarity of the maps as found by comparison with XFIT (Fig. 4.28). In the positionally fixed active site region, the similarity is even higher (86%), whereas in the C-terminal region, it is 81%. The last value is indicating a still relatively high agreement even for a strongly disordered part of the structure.

To obtain real-space map correlation coefficients separately for every residue of the model, an additional calculation was made with SFALL and OVERLAPMAP (CCP4 programs, see methods section). The comparison was first made without secondary conformations. The graph (Fig. 4.29) exhibits a better correlation for main chain atoms (blue curve) than for side chain atoms (red curve). For most of the residues, main chain correlation coefficients are over 70%, and for many even over 80%. The only negative exception is found for residues 221-224, where main chains and side chains both fit very poorly to the experimental map. This region has been mentioned before as flexible loop, and the very high B-values reflect the disagreement between the model and the weak electron desity as well.

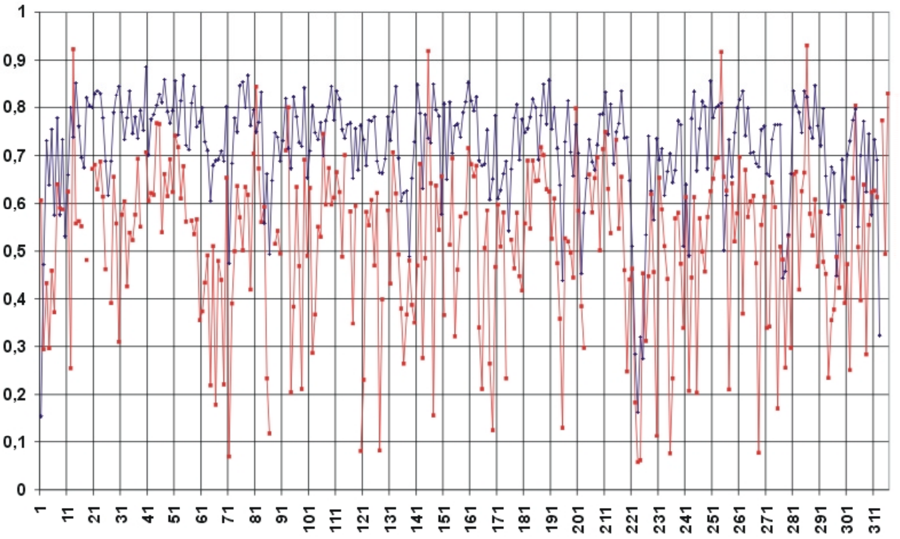
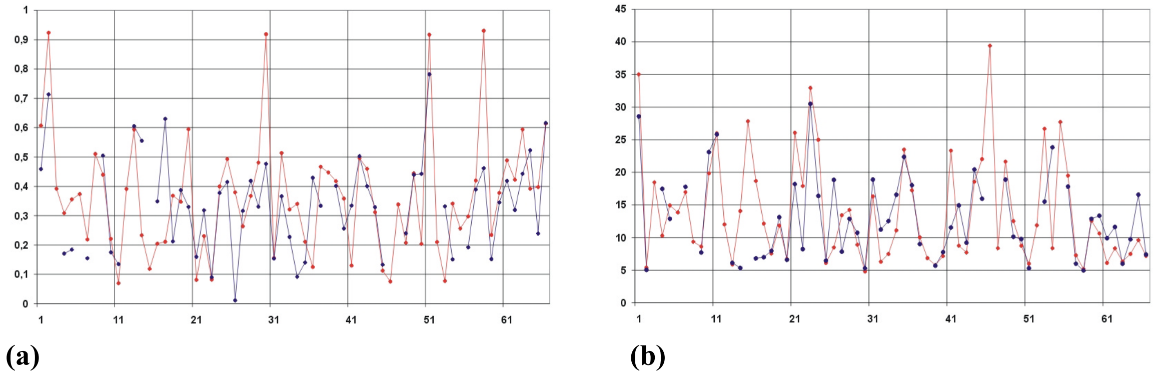


Fig. 4.29: CCP4 real space map correlation coefficients calculated for and plotted against single residues of hAR2. Main chain values are in blue and sidechain values in red. Secondary conformations were left out for this calculation.

Other residues, for which the side chain map CC is below 10%, are Glu70, Lys119, Glu126, Lys234 and Glu267. All of these are facing bulk-solvent regions of the structure and are therefore disordered. Looking systematically at the disordered parts of the structure, the correlation of the SHARP map to the model-calculated map was determined only for the residues with multiple side chain conformations, comparing individual curves for primary and secondary conformers (Fig. 4.30). Also the respective B-values were compared, because both quantities, B-values and correlation coefficients, are depending on the positional uncertainty of the atom positions.

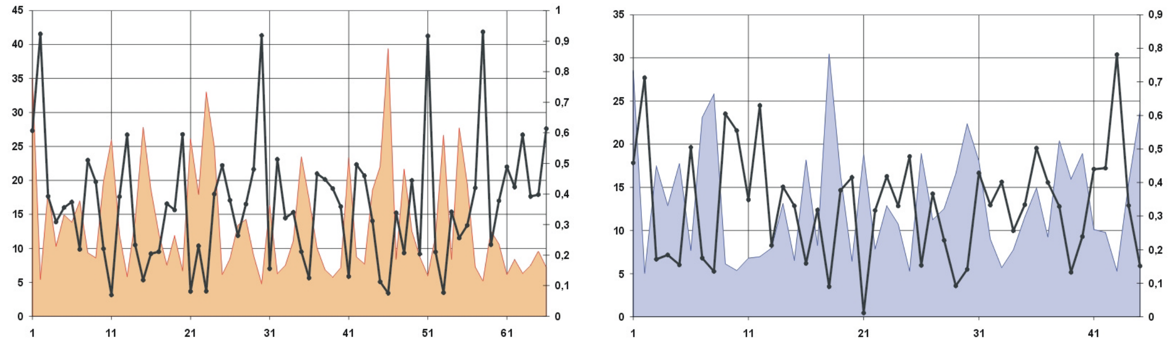


(a) **(b)**

Fig. 4.30: Comparison of modelled primary and secondary conformations and their agreement to experimental density. The red curves refer to primary conformations, the blue ones to secondary conformations. **(a)** The CCP4 real space map correlation coefficients. **(b)** the respective mean B-values for the residues.

Looking at the map correlation, it can be stated that for most of the residues, the agreement to primary conformations is better than to secondary conformations. This is plausible, because primary conformations are more visible in difference density maps, are usualy modelled more easily and accurately and are therefore more likely to be in the correct position. On the other hand, also experimental electron density taken for itself is better defined and less noisy if the underlying scatterers (i.e. the atoms in their true positions) have higher occupancies.

The positional fixation of atoms is decribed by anisotropic displacement parameters. It should be expected that residues with a low experimental map correlation have high mean B-values and vice versa. Fig. 4.31 confirms this expectation quite well.



(a) **(b)**

Fig. 4.31: The connection between B-values and map correlation coefficients, **(a)** for primary conformations of residues, **(b)** for secondary conformatons. The B-values are displayed as coloured areas, the CC values as black line graph.

Summing up sections 4.1.2 and 4.1.3, it can be stated that for very good anomalous data up to atomic resolution, the resulting phases and electron density maps are as well of extraordinary quality. The clear interpretability does not only allow an unproblematic autotracing, but also the determination or verification of structural details such as the modeling of multiple conformations (next section).

There are limitations, however, as far as hydrogen atom recognition is concerned. The experimental electron density map was thoroughly inspected for the appearance of hump-shaped expulsions of the otherwise spherical distribution around atoms. These humps corresponding to hydrogen atoms could only very rarely be found. The few significant 'humps' visible in a map of 0.7 σ contour could mostly be observed for polar hydrogen atoms involved in hydrogen bonds. This result was the same for a fully-resolved map and a map truncated to 1.1 Å (theoretically based on a better phase quality).

It should be kept in mind that the general assumption of delocalized electron density along hydrogen bonds (4 electrons altogether) is alternatively reasonable as explanation for the humps, which makes a statement a proof for hydrogen even less doubtful. Fig. 4.32 shows the only cases of humps associated to non-polar hydrogen atoms, however visible only at the reduced contour level of 0.7 σ.

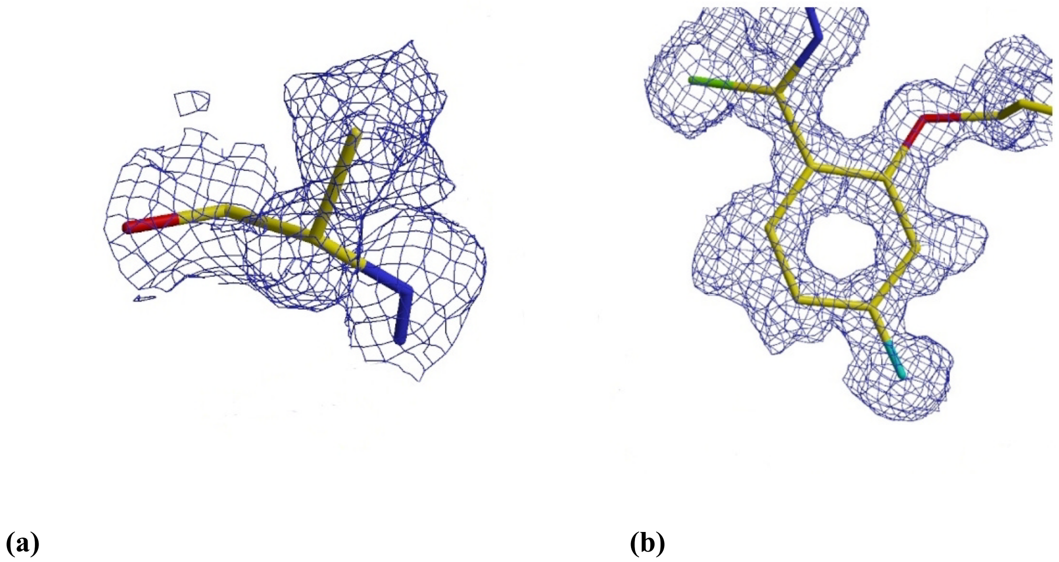
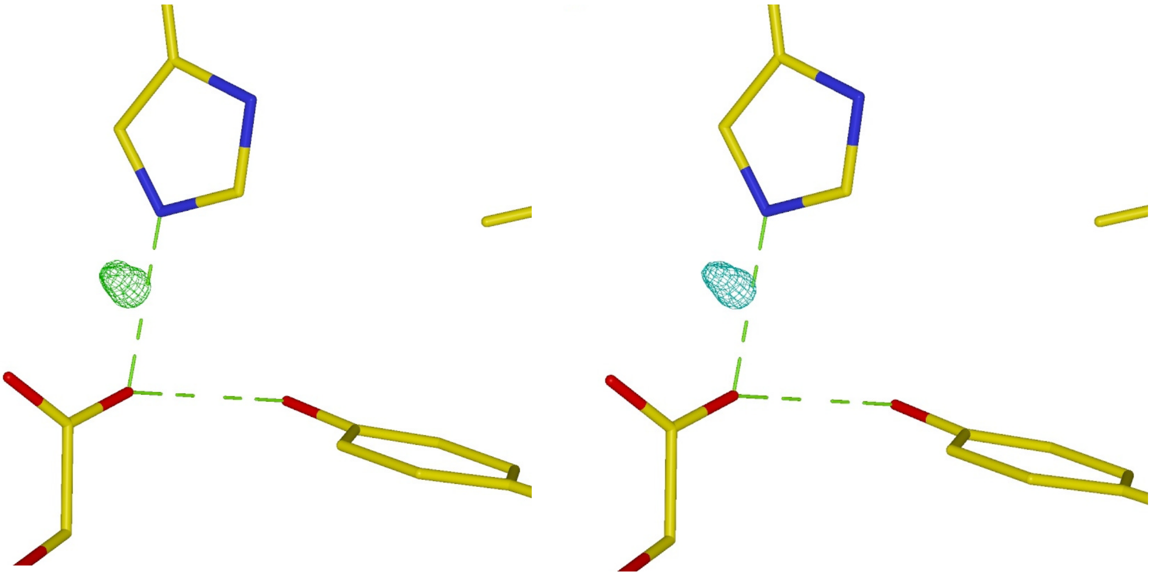


Fig. 4.32: Examples for hydrogen humps in the experimental FOM-weighted Fo,phi_{MAD} map, **(a)** an alanine residue, **(b)** one of the two phenyl rings of inhibitor IDD594. Both map images are show a contour level of 0.7 σ.

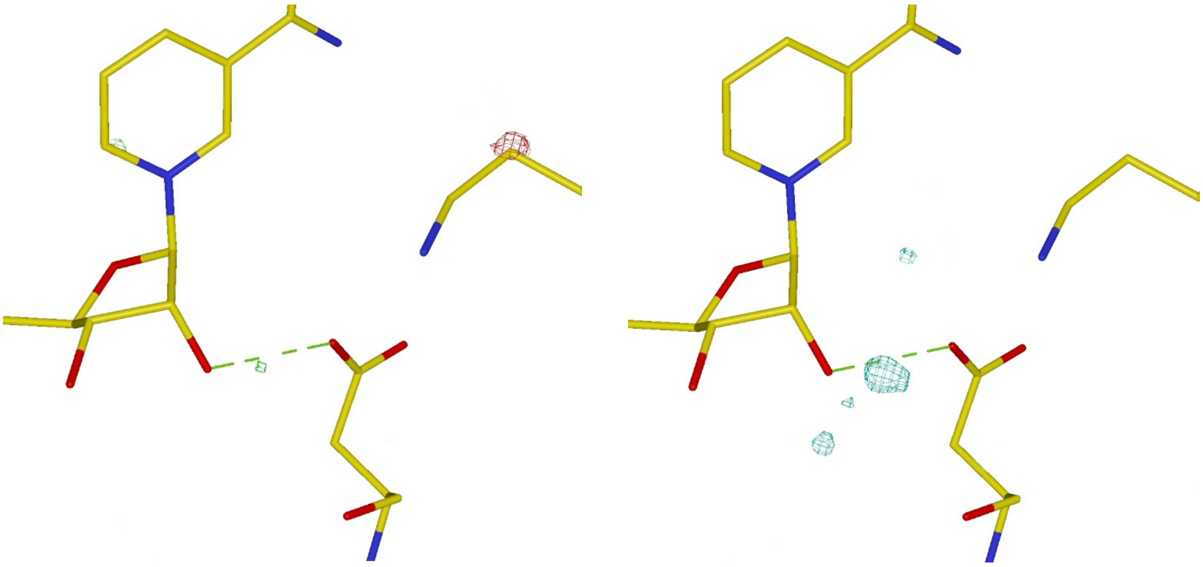
In general, clear spherical peaks in a difference electron density map facilitate the identification of missing model atoms like hydrogens. To combine this principal advantage of difference maps with the phase information from MAD, a experimental difference density map was used, containing SHARP amplitude and phase values combined with calculated SHELXL structure factor amplitudes. It was found that there are cases where experimentally phased difference peaks are larger (and higher) than those from conventional $F_o - F_c$ (phi_c) maps, but the improvement is in no other case as convincing as for Asp43 (Fig. 4.33). The case of His110 and Tyr48 shows virtually no difference between both map types (Fig. 4.34). Apparently, the experimental phases may lead to a somewhat more pronounced weak electron density, but also to a higher noise level, as some additional peaks in the Asp43 image indicate. There are a few hydrogen atoms like in the His110 case, which can easily be identified, but a majority of hydrogen atoms can either be identified very poorly (like for Asp 43, in that case more easily, if experimental phases are used) or not at all – like in the case of Tyr48.

On the whole, the exploitation of the experimentally phased electron density maps (F_o or F_o-F_c) does not lead to a noteable advantage compared to the refined F_o-F_c difference electron density map, as far as hydrogen atom identification is concerned.



(a) (b)

Fig. 4.33: The search for hydrogen atom peaks using two types of difference electron density maps, **(a)** conventional Fo – Fc, phic contoured in green and **(b)** Fo – Fc, phi$_{\mathrm{MAD}}$ contoured in cyan (right image), both contour levels are at 2.5 sigma. Diplayed are hydrogen bond interactions beween the carboxyl group of ligand IDD594 and the donor residues His110 and Tyr48.



(a) (b)

Fig. 4.34: Another example of difference density. The same map types and contour levels are used as for Fig. 4.33. Displayed is a hydrogen bond between one of the NADP+ sugar unit hydroxyl groups and the acceptor residue Asp43.

4.1.4 Classification of disorder using the experimental map

One main goal of the work on hAR was the systematic evaluation of modelled disorder using the bias-free experimental map obtained from sole SHARP phasing. The criterium for this was the presence or absence (as well as the interpretability) of experimental electron density at the primary and secondary conformer positions of the relevant protein residues. In this context, only the *visual* recognition of side chain shapes in the MAD map was relevant for the studies, not the numerical agreement to model-calculated maps. In other words, this part of the studies does not any more analyze the experimental phase / map quality (by refinement phase reference comparison) but, turning the point of view, the modeling is now validated *a posteriori* by the experimental map, which is assumed to represent the structural ”truth”.

For the 67 evaluated residues, both conformations were experimentally well recognized in 18 cases. For 27 residues only the major conformation and for 4 residues only the minor conformation was recognizable in the SHARP map. 21 residues remain, for which both positions of disorder are not or only poorly visible (see table 4.7) .

resnum	type	refined major [%]	refined minor [%]	cluster ID	exp. major consist.	exp. minor consist.	hygdogen bonds, partner residue	remark
Met1001	sc	65.4	34.6		O	O		atom type disorder possible, not modelled
Gln26	sc	65.9	- -		+	- -		part A only
Glu29	sc	53.6	(46.4)	I	+	+	(M)O: W: H-N Lys32 (m)O: H-O Ser127*(m)	
Arg40	bb	71.6	28.4		++	+		CA,C,O carbonyl split
His41	bb	71.6	28.4		++	+		N
Gln59	sc	76.9	(23.1)		+	-	[(M)N-H: W: H-N Arg63] (m)N-H: O=C Leu99	
Glu60	sc	68.9	- -	II	O	- -	(M)O: H-N Arg63(M)	major only, but reduced occ. from CD
Arg63	sc	68.9	(31.1)	II	++	-	(M)N-H: O Glu60(M) (m)N-H: O=C Asn171*	
Glu64	sc	68.9	- -	II	++	- -	(M)O: H-N Lys61 (M)O: W: O Glu60(M)	major only, but reduced occ. from CD
Val67	sc	58.7	41.3		+	O		
Lys68	sc	(50.8)	(49.2)	(III)	O	O	[(M)N-H: O Glu70(M)] only from exp.map	NOTE: swap A/B
Glu70	sc	(50.8)	(49.2)	(III)	O	-	[(M)O: H-N Lys68(M)] only from exp.map	
Glu71	sc	59.1	- -	IV	+	- -	[(M)O: W: O Glu70(M)] better in exp. map	major only, but reduced occ. from CG A: B 'bump' to Asp134*
His83	bb	59.1	40.9	IV	+	+		
Glu84	all	59.1	40.9	IV	O	+	(M)O: H-O Ser133(M) (M)O: W: O Asp134(M)	
Lys85	sc	70.2	- -		O	O*		* a 2nd is (better) only in exp. map
Gln93	sc	54.3	45.7		O	+	(M)N-H: O Ser97(M)	coupled to 97
Ser97	sc	54.3	45.7		++	+	(M)O: H-N Gln93(M)	coupled to 93
Asp102	sc	63.1	36.9		++	+	(M)O: W: H-N Lys234* [(M)O: H-N Lys234*]	diff. pos. for Lys in exp.map -> dir.bridge
Lys116	sc	65.3	34.7		+	- -		starting with CD CHECK
Pro117	sc	71.0	29.0		++	+		Pucker disorder (CG only)
Lys119	sc	(72.4)	(27.6)		-	O	pointing into BlkS.	
Glu120	sc	60.3	39.7		O	+	(m)O: H-N Lys239*(M)	occupancies don't fit
Glu126	sc	(53.6)	(46.4)	I	–	-	pointing into BlkS.	complementary A: B B: A 'bump' to Glu29*
Ser127	sc	53.6	46.4	I	+	O	(m)O-H: O Glu29*(m)	
Ser133	all	59.1	40.9	IV	++	++	(M)O-H: O Glu84(M) (M)O: H-O Thr235(M)	
Asp134	all	59.1	(40.9)	IV	+	-	(M)O: W: O Glu84(M)	A: B 'bump' to Glu71*

resnum	type	refined major [%]	refined minor [%]	cluster ID	exp. major consist.	exp. minor consist.	hygdogen bonds, partner residue	remark
Thr135	all	59.1	40.9	IV	O	+	(M)O-H: O Ser133(M)	
Asn136	all	59.1	40.9	IV	+	+	(m)N-H: W: O Asp139	
Ile137	all	59.1	40.9	IV	+	O		
Glu146	sc	53.7	46.3	VI	-	O	(m)O: W: O Glu150 ---- (M)O: H-N Arg293*(M)	— O: H-N Arg293*(m) NOTE: swap A/B
Leu152	sc	67.6	32.4		+	O		
Lys154	sc	57.8	42.2		+	O	(M)N-H: W: O=C Gly151	minor position does not differ signific.
Asn162	bb	60.6	39.4	V	+	+		
His163	all	60.6	(39.4)	V	+	-	on blk solv surface	A: B 'bump' to Phe315
Leu164	all	(60.6)	(39.4)	V	-	- -		
Met168	sc	(41.8)	(35.6)		+	-		3rd conf. probable, atom type diso. poss. both not modelled
Ile169	sc	64.8	35.2		+	O		
Lys178	sc	54.3	- -		++	- -		only A part
Pro179	sc	58.3	41.7		O	O		Pucker disorder (CG only)
Glu193	sc	53.4	46.6		++	+	(M)O: H-O Thr191(M) (m)O: W: O=C Lys21*	
Lys194	sc	(60.6)	(39.4)	V	O	-	(M)N-H: W: H-O Thr265* (m)N-H: O Gln26*	
Gln197	sc	65.9	- -		++	- -		only A part
Arg217	sc	65.5	34.5		+	O	(M/m)N-H: O Asp224 (M)N-H: O Ser214	
Glu229	sc	71.2	28.8		O	O		
Lys234	sc	53.7	- -		+	- -		only A part
Lys239	sc	58.6	- -		+	- -	(M)N-H: O Asp284	major only, but reduced occ. from CD
Asn241	sc	62.5	37.5		O	O	pointing into BlkS.	
Lys242	sc	73.5	26.5		+	O	(M)N-H: W2469 (m)N-H: O Asp277	Parallel shift
Thr244	sc	62.2	21.8		+	+		3rd conformation with occ. = 29.0 %
Asn256	sc	74.1	- -		++	- -		only A part
Glu267	sc	(60.4)	(39.6)		-	-	pointing into BlkS.	
Glu271	sc	(65.5)	(34.5)		+	-	(M)O: H-N NAD (318) (M)O: H-O Thr243	
Asp277	sc	73.2	- -		++	- -		only A part
Glu279	sc	(57.2)	(42.1)		-	-	(M)O: H-N Arg250	

resnum	type	refined major [%]	refined minor [%]	cluster ID	exp. major consist.	exp. minor consist.	hygdogen bonds, partner residue	remark
Ser282	sc	42.4	34.4		++	++		3rd conformation with occ. = 24.9 %
Arg293	sc	(53.7)	(46.3)	VI	-	+	(M)N-H: O Glu146(M) (m)N-H: O Glu150 ----	— O: W: O Glu146(m)
Trp295	all	56.7	43.3		++	+		
Arg296	all	56.7	43.3		+	O		
Val297	all	56.7	(43.3)		+	-		
Cys298	all	56.7	43.3		+	+		
Ala299	all	56.7	43.3		++	+		
Leu300	bb	56.7	43.3		++	+		
Ser305	sc	73.8	26.2		++	+	(M)O-H: W: O Asn129	
Phe311	bb	63.6	36.4		+	O		
His312	bb	63.6	36.4		O	O		
Glu313	bb	- -	(36.4)		O	- -		only N, only minor

Table 4.7: The disordered residues of hAR2. resnum corresponds to residue number, type to type of disorder (bb = all backbone (i.e. N-CA-C-O) atoms, sc = all side chain (i.e. not backbone), all = all atoms of one residue, major/minor occupancy as refined, in brackets if weak refinement map correlation, missing (- -) if only primary conformation modelled.
Experimental map correlation qualifiers are used as follows: - - no or almost no correlation, - weak only, O quite acceptable, + satisfying, ++ very good. H-bonds: * indicates symmetry equivalent

Looking at Table 4.7, a coincident agreement of the model to the refinement maps and to the experimental maps, respectively, is found. Protein residue conformations that were difficult to model because of missing difference density peaks, are in many cases also lacking density in the experimental map. Disordered side chains exhibiting negative density even for the primary conformer, thus indicating a non-optimal placement or refinement, show in general a poor experimental electron density at both conformer positions. It seems likely that those problematic cases can be explained by multiple disorder, where the assumption of only two dominating conformations does not hold. Since no density modification was applied to the SHARP map, missing experimental density at low occupied side chains *cannot* be explained by artificial downweighting effects.

“Delocalized” side chains are expected at the surface of the protein, lacking coordination partners (other sidechains or positionally fixed water molecules). Therefore, it is not surprising that most of the disorded residues, especially those with problematic modeling, belong to typical polar or charged amino acid types, usually found at the protein surface, exposed to water. The most frequently found residues in the hAR2 case are glutamate and lysin (Table 4.8), being amino acids with especially long side chains. As far as side chain torsion angles are concerned, glutamate has got three degrees of conformational freedom, and lysin even four. Therefore the combinatorial probability of conformational disorder is very high, also explaining these results. Arginin is an amino acid with even five degrees of freedom. The reason for the less frequent appearance of Arginin in the list of disordered residues is most likely the smaller fraction in total (11 Arg in the hAR model, 3.5%, while there are 22 Glu, 7.0%, and 25 Lys, 8.0%).

	category	number of resid.	favourite residue types (number)	mean occ. 1st conf.	mean occ 2nd. conf.	mean B 1st conf.	mean B 2nd conf
I	both sidechains	18	Ser (4), Glu (2)	59%	38%	9.0	10.0
II	one sidechain	31	Glu (5), Lys (5)	65%	28%	12.7	12.5
III	no sidechain	18	Glu (7), Lys (5)	63%	35%	22.1	19.4
	total	67	Glu (14), Lys (10)	63%	33%	13.4	13.0

Table 4.8: Summary of categories for modelled disorder with respect to the positional side chain visibility in the experimentally phased map. B-values are given as isotropic equivalents of the ADPs. The total mean B-values were calculated from all relevant atoms seperately. Note that residues with backbone-disorder only were not included in these statistics. Occupancies do not sum up to unity, because two tertiary conformations were ignored for type I, and several occupancies are 0% for missing secondary sidechain parts (type II only).

In fact, almost all disordered residues are on the protein surface or very close to it. However, there is an important difference between residues of types I and II, and type III residues on the other hand. While the former are located at interfaces to symmetry equivalent hAR2 macromolecules, are half-buried in the protein surface or belong to well ordered water networks, the latter are mostly exposed to "empty" regions of the structure, *i.e.* the bulk solvent region. Therefore, all side chains of type I / II residues have several defined interactions to other protein residues or positionally fixed water molecules. The type III sidechains however are lacking those contacts. This is likely the most significant explanation why those residues are prone to build multiple conformations, for which the two modelled disorder components are only poorly fitting descriptions.

Looking more closely at the average refined occupancies and B-values of both conformer ensembles, a significant difference of B-values between types I / II and III is striking. The comparably low mean B-values of the first two point out that, in general, both conformers of these types were refined in reliable positions and with realistic occupancies, even for many of the secondary disorder components. The poor visibility of the secondary side chains of type II in the MAD map is in some cases (if the refined B values are comparably low for both conformers, like for Gln59) just due to the fact that their occupancies are very low, and the weak density is just the result of this fact. Still, it can be assumed that these weak secondary conformers are completely correct, *i.e.* that there is probably no third conformation present. In other cases, like Asp134 or Gln197, the B-values of the minor conformer are much higher than ones for the major conformer. In these cases, a well-placed first conformation can be assumed, whereas the second may be further disordered and is insufficiently modelled by just one minor conformation. For eight residues of type II, this circumstance has even prevented the modeling of a second sidechain (occupancy values are missing in the respective column of Table 4.7). Five cases exist, where obviously the occupancies were wrongly assigned, which is reflected by the fact that the experimental visiblitity is better and the B-values are lower for the minor conformer. This result is particularly interesting because there is a disagreement between the conventional refinement with difference maps and observations in the experimental map.

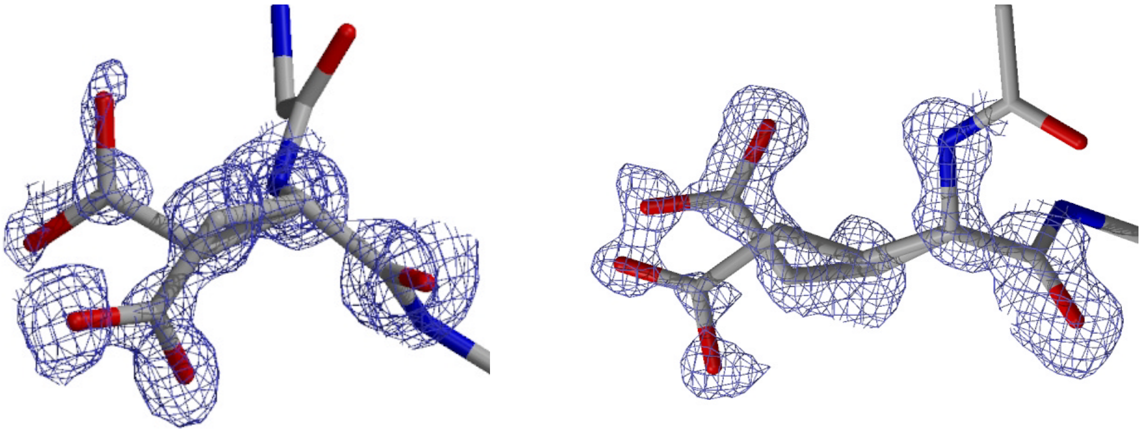

(a) (b)

Fig. 4.35: Examples for twofold side chain disorder with good visibility in the experimental map for both conformations. **(a)**: Residue Asp102 (Occupancies are 63% : 37%). **(b)**: Residue Glu193 (Occupancies are 53% : 47%).

Asp102 and Glu193 are examples for the first type of side chain double-conformer residues, where both conformations were easily modelled due to difference density maps and both are recognizable in the MAD map (Fig. 4.35). They are both located in surface regions facing highly ordered water networks. The primary conformer of Glu193 is additionally bending back towards the protein and interacts with Thr191, whereas the secondary side chain is close to Lys21 (belonging to a symmetry equivalent molecule). They are indirectly connected via H-bonds with a water molecule. The interpretability of the MAD map is so good for these two residues, that even the occupancy differences of 26% (Asp102) and 6% (Glu193) can be nicely recognized and therefore confirmed.

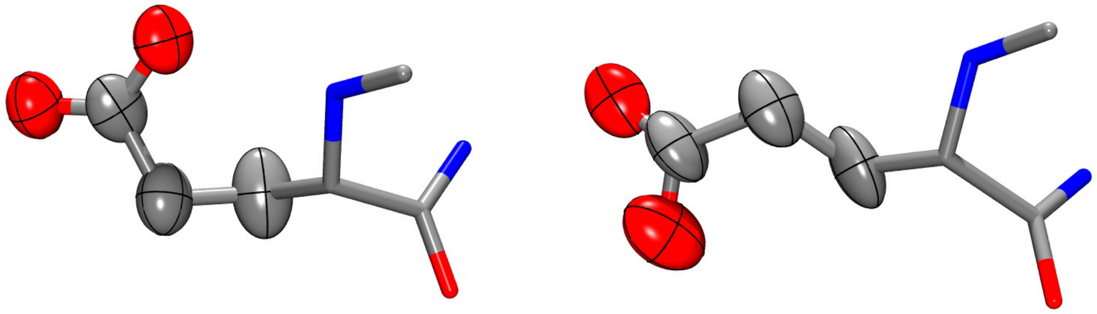

(a) (b)

Fig. 4.36 Thermal (50% probability) ellipsoids representing the ADPs for the Glu193 side chain atoms. **(a)** the primary, **(b)** the secondary conformation is displayed.

Glu193 has a mean B value of 7.14 Å^2 (isotropic equivalent) for the side chain atoms of the primary and 7.76 Å^2 for the sidechain atoms of the secondary conformation. These displacements are very low for a disordered residue (the overall mean sidechain B value is 13.2 Å^2 for all disordered residues and 9.5 Å^2 for all members of category I). This fact, underlining the particularly good modeling of Glu193, is visualized in Fig. 4.36. The mean sidechain atom B values are 7.54 Å^2 and 7.69 Å^2 for the Asp102 conformers.

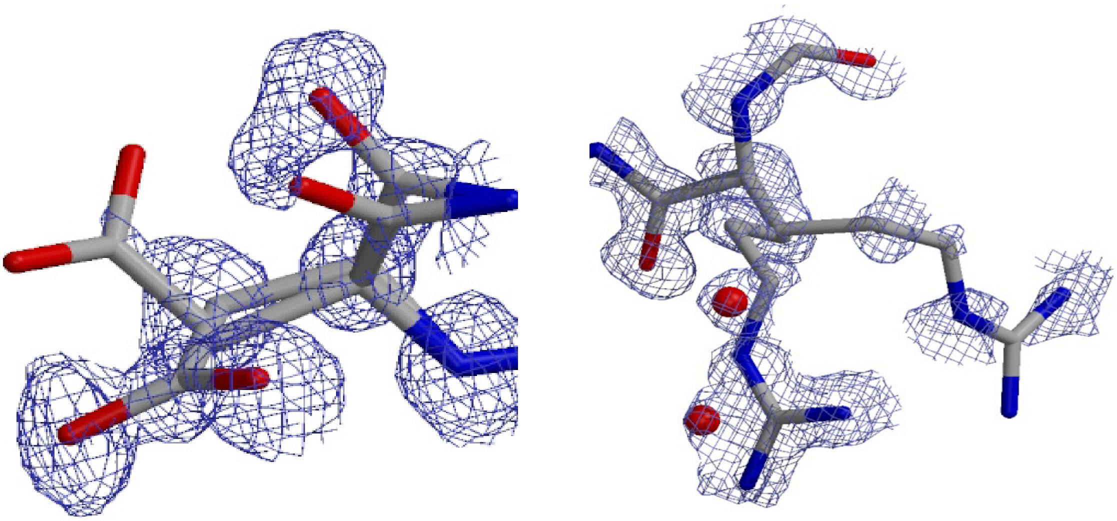

(a) (b)

Fig. 4.37: Examples for twofold side chain disorder with good visibility in the experimental map only for the primary conformation. **(a):** Asp134 (Occupancies are 59% : 41%). **(b):** Arg293 (Occupancies are 54% : 46%).

Asp134 and Arg293 are examples for the second type of residues modelled with twofold sidechain disorder. Their modelled positions exhibit good agreement to the experimental electron density map shape only for the primary conformer. For the secondary side chain, MAD electron density is present only at some atom positions (Fig.4.37).

Both residues are located at interfaces to symmetry-equivalent hAR2 molecules or other parts of the same molecule, so they both participate in disorder clusters with common free occupancy variables. The primary conformer of Asp134 interacts indirectly with Glu84, bridged by a water molecule. For the secondary side chain, however, no contacts were found. Arg293 has got coordination partners for both conformers, Glu146 and Glu150 respectively.

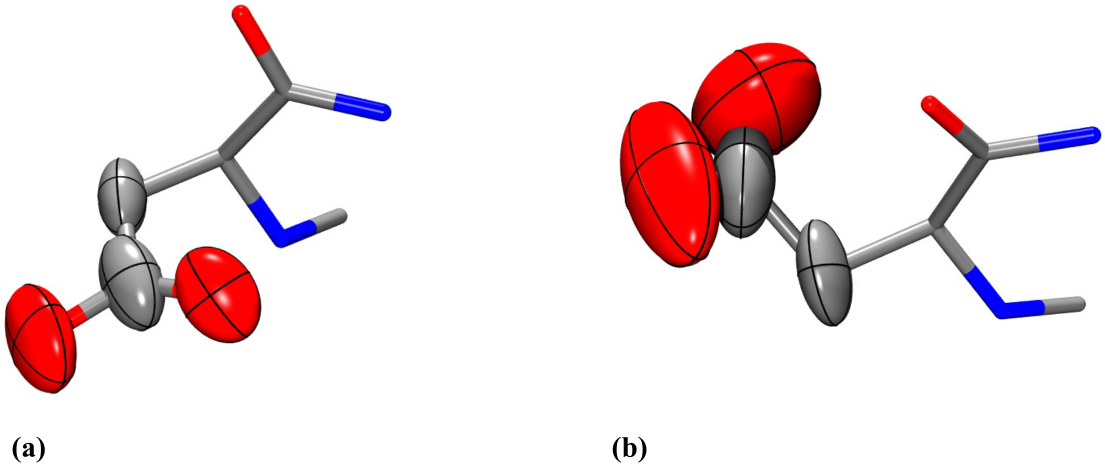


Fig. 4.38: Thermal (50% probability) ellipsoids representing the ADPs for Asp134 side chain atoms. **(a)** the primary, **(b)** the secondary conformation is displayed.

Asp134 has a mean B value of 8.46 Å^2 (isotropic equivalent) for the atoms of the primary and 18.85 Å^2 for the atoms of the secondary conformation. In case of Arg293, the values are 12.58 Å^2 and 12.86 Å^2. In case of Asp134, the secondary sidechain modeling and occupancy refinement apparently fit the real situation worse than for Arg293B, where also the experimental electron density is somewhat better. Additional disorder positions of the secondary sidechain (*i.e.* the existence of multiple conformations) would explain the high B-values and the weak MAD density around Asp134B – this explanation does also agree with the missing contacts.

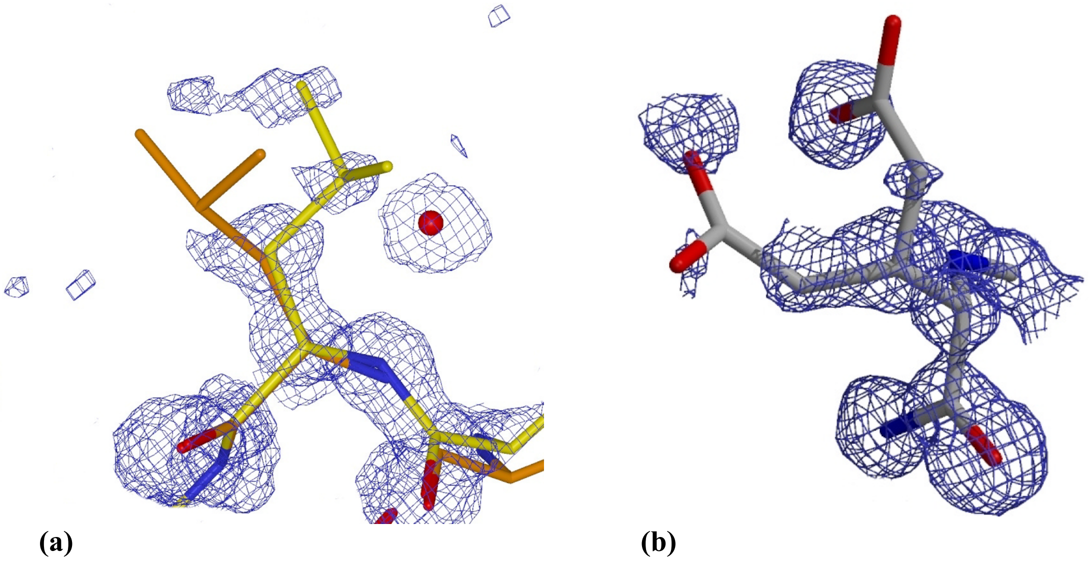


Fig. 4.39: Examples for twofold side chain disorder with bad visibility in the experimental map for both conformations. **(a):** Leu164 (Occupancies are 61% : 39%). **(b):** Glu267 (Occupancies are 60% : 40%).

Leu164 and Glu267 are examples for the third type of side chain double-conformer residues, where both conformations were modelled with difficulties due to unclear difference density maps – neither are they sufficiently recognizable in the MAD map (Fig. 4.39). Both Leu164 and Glu267 are pointing into solvent water regions, lacking water contacts. Leucine, of course, cannot build hydrogen bonds. Therefore it is very surprising to find this lipophilic residue at the hAR2 surface, anyway.
The occupancy difference between major and minor conformation is comparable in the two cases. The sidechain atom mean B values are 23,51 Å^2 : 22,35 Å^2 for Leu164 and 26,69 Å^2: 15,47 Å^2 for Glu267. The last value indicates that the secondary conformation of Glu267 should maybe have a higher occupancy.

The subject of wrong major/minor conformation assignments due to wrongly refined occupancies leads to the interesting cases, where the agreement to the MAD map is better for the secondary conformation – as found for Glu84, Gln93, Glu120 and Thr134. Fig. 4.40 shows that for residue Gln93, the σ_A and F_o-F_c type maps of the refinement do not disagree to the given model in such a way that a occupancy correction seems obvious. All primary side chain atoms fit the σ_A shape, and the rather small negative difference density area covers neither Cδ nor Nε precisely. The MAD map, however, points out much more clearly that the so-called major confirmation is actually weaker than the minor one, lacking electron density both at the oxygen and nitrogen atom of side chain Gln93A.

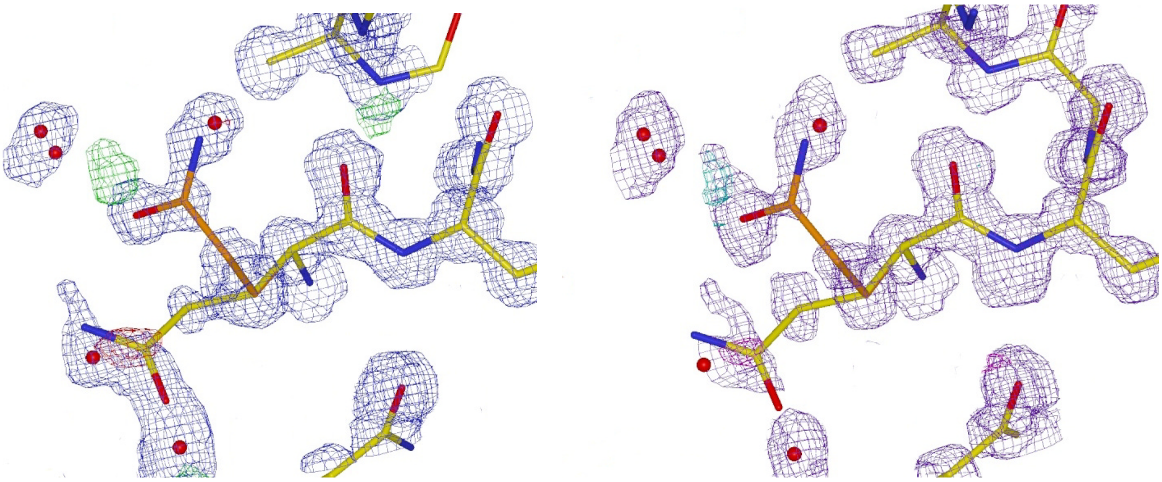

(a) (b)

Fig. 4.40: The double conformation of residue Gln93, modelled with 54:46% occupancies. **(a)**: σ_A map in blue (1 σ level) and Fo-Fc map in green (+3 σ) and red (-3 σ). **(b)**: Fom-weighted Fo, phi$_{MAD}$ map in blue-violet (1 σ level) and Fo-Fc, phiMAD map in cyan (+3 σ) and magenta (-3 σ).

Side chain atom mean B-values are 18.68 Å^2 for Gln93A and 6.80 Å^2 for Gln93B. Of course, this is a clear sign for either wrong placement of the primary conformer or wrongly refined occupancies. Compared to using MAD map, there is a practical disadvantage of using B-values in order to check the occupancy correctness – the B-values are more likely to be overseen, because they can only be evaluated after the next refinement, whereas the MAD map can be used *a-priori*. During the refinement of hAR2, the MAD map was not used at all, but on the other hand, more attention was given to the σ_A and difference electron density maps than to the refined B-values.

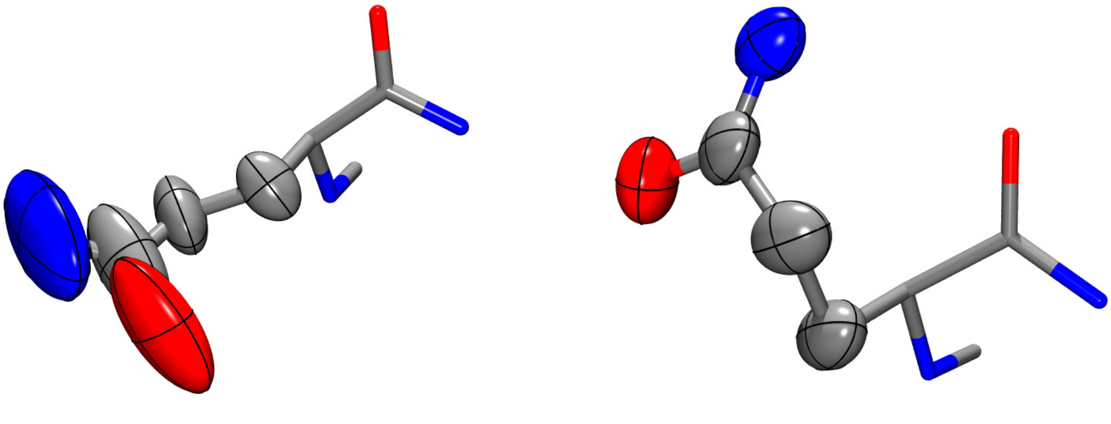


Fig. 4.41: Thermal ellipsoids representing ADPs for side chain atoms of the **(a)** primary and **(b)** secondary conformation of residue Gln93.

All that was said for Gln93 is also true for the other three residues. The mean B-values are 14.08 Å^2: 5.34 Å^2 for Glu84, 17.91 Å^2 : 8.22 Å^2 for Glu120 and 13.42 Å^2 : 7.84 Å^2 for Thr135. At least for residues Glu84 and Thr135, the wrong "individual" values could be explained by the fact that they belong to a large cluster and have been "forced" to fit one common free occupancy variable. The principal questions is, how equal the occupancies of clusters really are (theoretically, all cluster members should have exactly the same occupancy). If ADPs clearly propose another occupancy distribution, the residue assignments to the cluster have to be revisited.

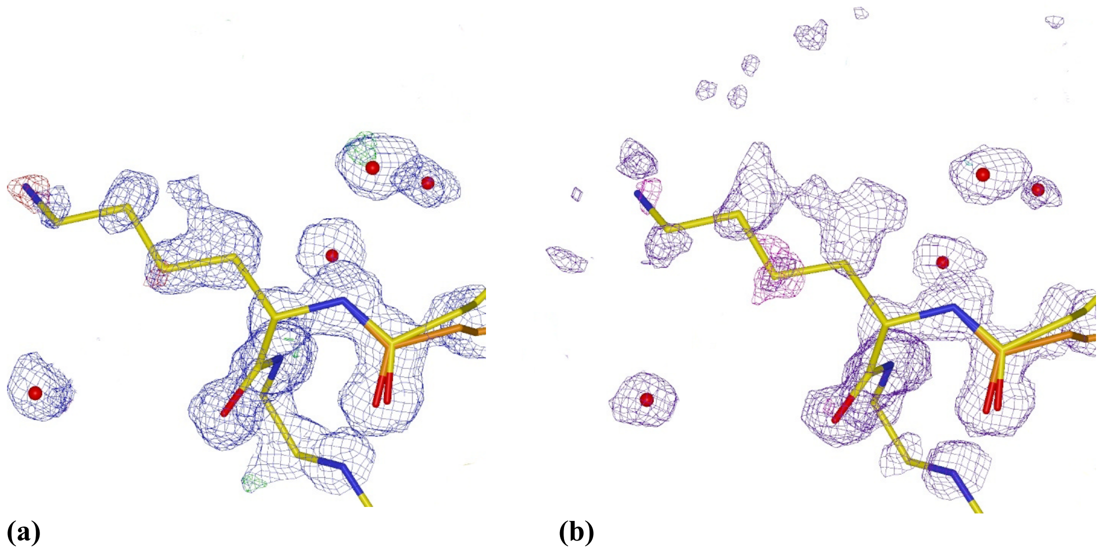


Fig. 4.42: Model of residue Lys85 with model-calculated maps **(a)** – σ_A, level one σ, blue, and Fo-Fc, +/- 3 σ, green / red) and experimentally phased maps **(b)** – Fom-weighted Fo, one σ, blue-violet, and Fo-Fc, +/- 3 σ, cyan/magenta)

Lys85 could only be modelled in one conformation with reduced occupancy due to the model-phased refinement map. The experimental F_o map reveals an alternative conformation (Fig. 4.42) with a more recognizable density. This special result is emphasized here, because it is a particularly clear example of disagreement between σ_A and MAD map, apparently to be explained by model bias. Atom Cγ seems to fit both refinement map types well – there is a well-shaped σ_A peak and only a small negative difference density. In the experimentally-phased maps, Fo density is much weaker than σ_A and the F_o-F_c difference map exhibits much more negative density at the Cγ position.

4.2 The Application of SITCOM

4.2.1 The verification of program functionality with test structures

The basic purpose of the test structures used with SITCOM was to check the technical functionality of the program, *i.e.* to ensure that all algorithms work for all principal space group types - non-polar (possibly high-symmetric) ones, polar ones and the P1 case. The five structures chosen for the test suite covered this range of spacegroups, and they differed also in the number of (selenium) sites, ranging from 3 to 59.

working name	space group	d deriv.	Se sites	PDB native	HA native[1]	d native[1]	reference
ModE	$P2_12_12$	1.75 Å	2 x 3	1B9M	Se	1.75 Å	Hall et *al.* 1999
JIA	$C222_1$	2.50 Å	2 x 4	1C8U	S	1.90 Å	Li et *al.* 2000
RRF	$P4_32_12$	2.80 Å	1 x 3	1DD5	S	2.55 Å	Selmer et *al.* 1999
TransH	$P2_1$	2.00 Å	4 x 15 [2]	1F8G	Se	2.00 Å	Buckley et *al.* 2000
Cyan	P1	1.65 Å	10 x 4	1DW9	Se	1.65 Å	Walsh et *al.* 2000

Table 4.9: Overview of the five test structures used to verify the SITCOM functionality. HA native is the “heavy atom” type of the (pseudo-) native data. [1] Only for JIA and RRF refined native structures were available, the other PDB files contained refined derivative structures [2] TransH is lacking one site for one monomer, so it contains 59 selenium sites.

As many of the test results were useful for program debugging, but are only of technical interest, they will not be discussed here. Transhydrogenase B is an exception, because it was used in detailed studies on dataset-related substructure accuracy.

Depending on the presence of refined heavy atom positions from a completed structure determination, SITCOM may be used for a-posteriori analysis of subtructure solutions. If a PDB file containing the refined positions is read by SITCOM, the program will automatically compare all solution trials with these “reference sites” instead of cross-comparing the experimental substructures.

4.2.2 Studies on Transhydrogenase B

Transhydrogenase B (THB, Buckley et *al.* 2000) is a tetramer of 384 amino acid residues per monomer. It contains 59 selenium sites, 15 per monomer except for one with only 14 sites. THB crystallises in the polar monoclinic spacegroup $P2_1$. The three MAD datasets have a limiting solution of 2.0 A. Also the refined model used as reference is based on a selenium derivative structure, which had been refinened against 2.0 A data.

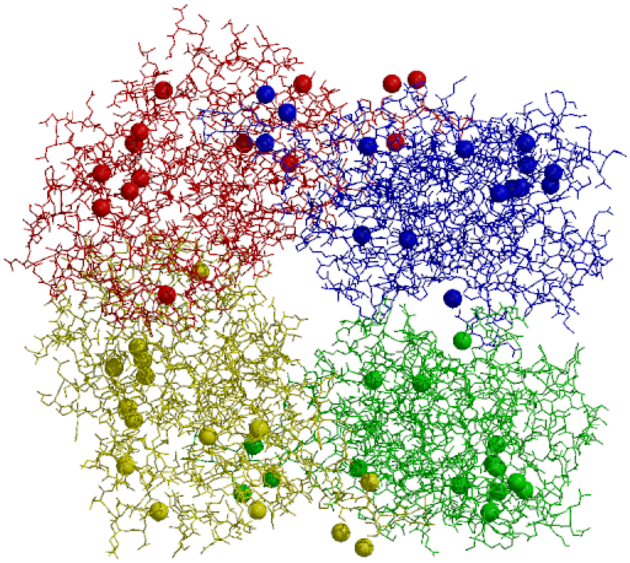


Fig. 4.43 The structure of THB with all bonds displayed in wireframe style. Special colours are given for each chain. The selenium atom positions are highlighted as spheres.

Because of its many selenium sites, not all of which are positionally well fixed and easily found, THB is very well suited to serve as study object to investigate how the choice of different MAD data subsets affects the substructure quality. A substructure solution can be called accurate if many trial sites are found that correspond to refined atoms (within a certain distance limit), and if the mean distance of all corresponding sites to their reference positions is small.

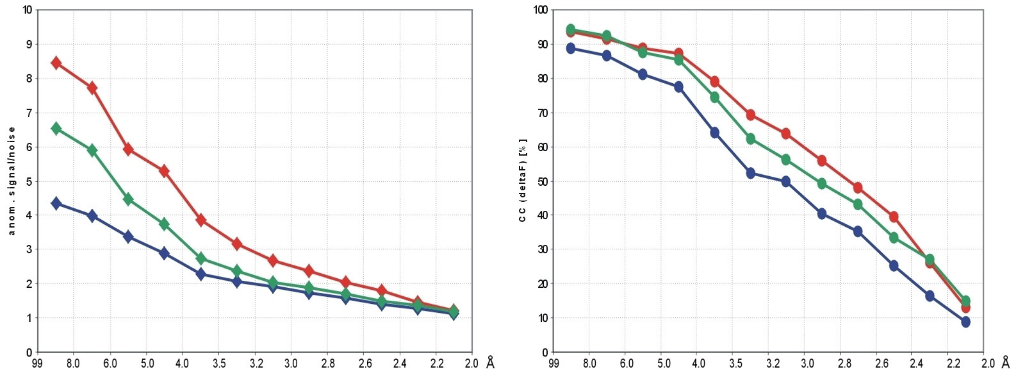


Fig. 4.44: Data quality of the single wavelength data subsets. Left: Anomalous signal-to-noise ratio, $<\Delta F/\sigma(\Delta F)>$ in resolution bins. Peak data values are plotted in red, high-energy remote data values in green and inflection point data values in blue. Right: Inter-wavelength CC (ΔF_i, ΔF_j) for each wavelength combination in resolution bins. The red graph is for (peak, remote) the green one for (peak, inflection) and the blue one for (remote, inflection).

The analysis of the anomalous MAD datasets with XPREP (Fig. 4.44) reveals a normal trend of anomalous signal-to-noise ratio, with the peak data having the highest intensity over sigma, followed by high energy remote data and inflection point data being weakest. The type of decrease is rather linear, although steeper in the lower half of the resolution range.

Peak and high-energy remote data correlate best, inflection point and high-energy remote data worst. Again, the decrease of correlation is quite linear for all combinations, but this time it is steeper at the high resolution edge. This as well as the order of data subset correlation curves can be explained by the fact that the correlation coefficient depends on the signal-to-noise ratio.

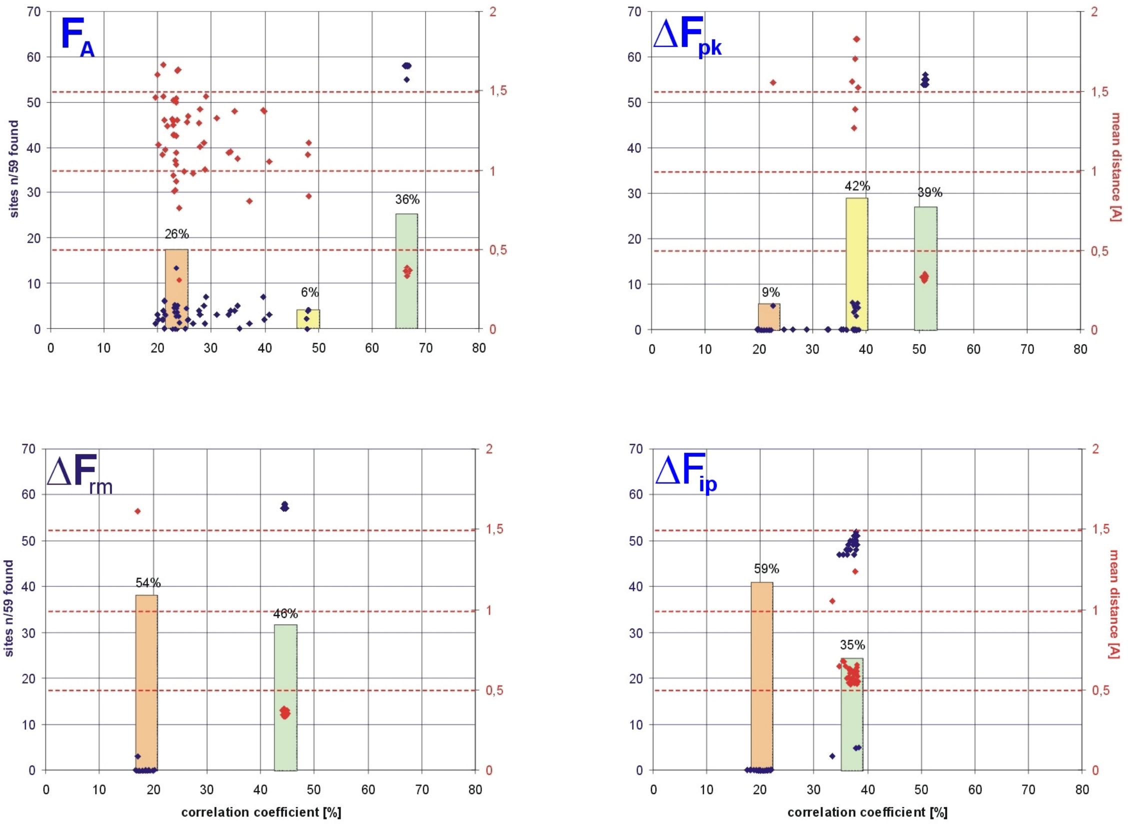


Fig. 4.45: Overview of substructure solutions at 3.0 Å for the different wavelength data subsets (pk = peak, rm = high-energy remote, ip = inflection point). Each scatterplot displays 100 SHELXD solutions. Both the numbers of refined sites found (blue) and the mean distances between the trial/reference pairs (red) are plotted against the correlation coefficients. The bars indicate clusters of solutions within a close CC range. The number of solutions per cluster relative to all 100 is given as percentage.

Looking at the SHELXD output for data subsets of four types (F_A and the anomalous differences for the three experimental wavelengths, Fig. 4.45), clusters of solutions with very similar correlation coefficients (*i.e.* mostly bimodal distributions) can be observed. Only the clusters with the highest respective correlation coefficients contain 'good' solutions, for which most of the sites correspond to true selenium positions – independently of the absolute $CC(E_{obs},E_{calc})$ values. The $F_A$data solutions have the highest correlation coefficients and are less clustered than the rest – the CC values as well as the mean distance values are widely scattered.

Only one solution of the top cluster contains less than 58 true Se positions. For ΔF_{rm}, the largest fraction of solutions is in the top cluster, but most of them comprise only 57 corresponding sites. ΔF_{pk} produces less true positions, but here the <d>-values are smallest. The ΔF_{ip} substructures are much less accurate than the rest.

Analyzing the best solutions (with highest correlation coefficients) for each data subset and resolution limit (table 4.10), the combination of F_A data and a 3.5 Å threshold leads to the best absolute result, which is a solution with 58 true selenium sites, having a mean distance of 0.35 Å to the refined positions. For the single wavelength data subsets, a correlation between the type of subset and the resolution limit can be observed: the best result for peak data is obtained with a resolution limit of 2.5 Å, for the high-energy remote data, the threshold is 3.0 Å and for the inflection point data it is 3.5 Å. One can conclude that the better the anomalous signal-to-noise ratio (pk > hrm > ip, see Fig. 4.44), the more will the substructure quality benefit from a higher resolution threshold and thus a larger amount of data. In other words, as long as there is a significant intensity over sigma, as many data as possible should be used for the substructure solution, because of the information gain related to high resolution.

d_{min}	**F_A MAD**			**ΔF peak**			**ΔF high-en. remote**			**ΔF inflection point**		
[Å]	CC_1 [%]	n/59	<d> [Å]	CC_1 [%]	n/59	<d> [Å]	CC_1 [%]	n/59	<d> [Å]	CC_1 [%]	n/59	<d> [Å]
2.0	35.9	52	0.44	37.4	54	0.29	26.4	54	0.43	11.6	0	--
2.5	55.1	56	0.36	**46.4**	**57**	**0.27**	37.8	55	0.32	17.2	0	--
3.0	66.9	58	0.36	51.3	55	0.32	**44.7**	**57**	**0.37**	38.1	49	0.56
3.5	**71.7**	**58**	**0.35**	53.9	56	0.39	49.2	57	0.40	**42.1**	**52**	**0.63**
4.0	74.4	58	0.42	32.5	0	--	29.5	0	--	38.9	37	0.76

Table 4.10 Overview of best solutions (according to highest CC) for each wavelength data subset and limiting resolution. For each solution, also the number of refined sites found and the mean distance between the trial/reference pairs are given. The best resolution limits for each set are pointed out by bold text.

Surprisingly, at 3.0 Å, high-energy remote data leads to better *absolute* results than peak data, although peak data has the larger intensity-over-sigma throughout the whole resolution range. Presumably, the data errors themselves have a greater influence on substructure accuracy than just the ratio between anomalous intensity and error. In this respect, the remote data should be most precise, as the anomalous signal (f'') does not significantly change with wavelength fluctuations.

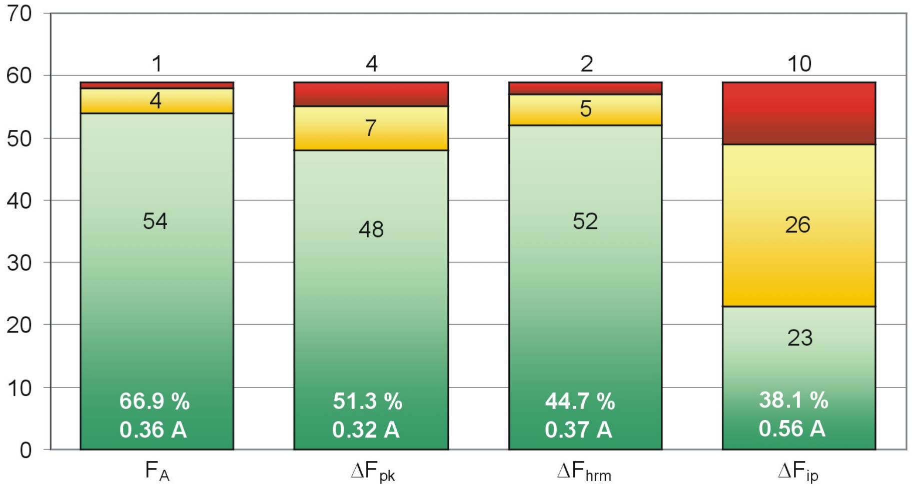


Fig. 4.46: Correspondance between the 59 refined atom positions and the trial sites of the best data subset solutions at 3.0 Å (CC values and mean distances are given, see also line 5 in Table 4.10): Green bars are used for continuous trial sites with a refined partner, yellow bars for additional non-continuous sites (i.e. after the first no-partner site), and red bars for sites without any partner.

In general F_A and ΔF_{rm} are superior to ΔF_{pk} with respect to the solution with highest CC (CC_1) at 3.0 Å (Fig. 4.46), whereas ΔF_{ip} is much worse that the rest: only 49 sites correspond to the refinement, and site #24 is already the first site for which no partner is found. It has to be emphasized that the CC_1 solutions are not necessarily the ones with most selenium positions found or with lowest mean distances (<d>). For ΔF_{hrm}, there are seven solutions with 58 Se found - one more than CC_1. The <d>-value of each CC_1 is comparable to the mean <d>-value of the respective top cluster (except for ΔF_{ip}), but never equals the $<d>_{min}$ -value.

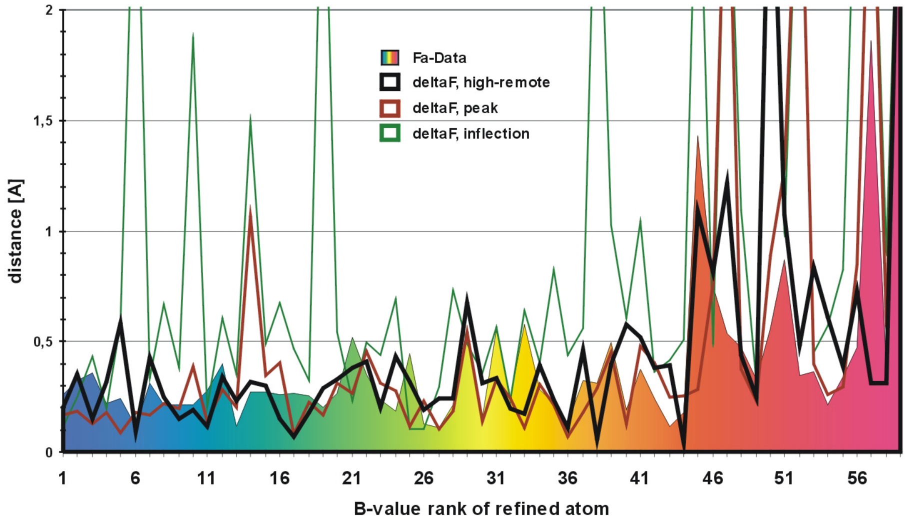


Fig. 4.47: Positional deviations of the 3.0 Å (best solution) single trial sites plotted against the corresponding refinement atom number . The refined atoms are sorted from low to high B-value. The colours are used as explained by the legend: the filled, rainbow coloured graph (B-value temperature colour scheme from cold to hot) is for Fa data, the black one for high-energy remote, the brown one for peak and the green one for inflection point data.

As pointed out in Fig. 4.47, the locational accuracy of the experimental sites also depends on the positional displacement of the respective selenium atom, as expressed by the refined B-values. If a selenium atom has got a high temperature factor (possibly belonging to a disordered methionine side chain) it can not precisely be determined in a substrucure solution and therefore also the distance to the reference position is high.

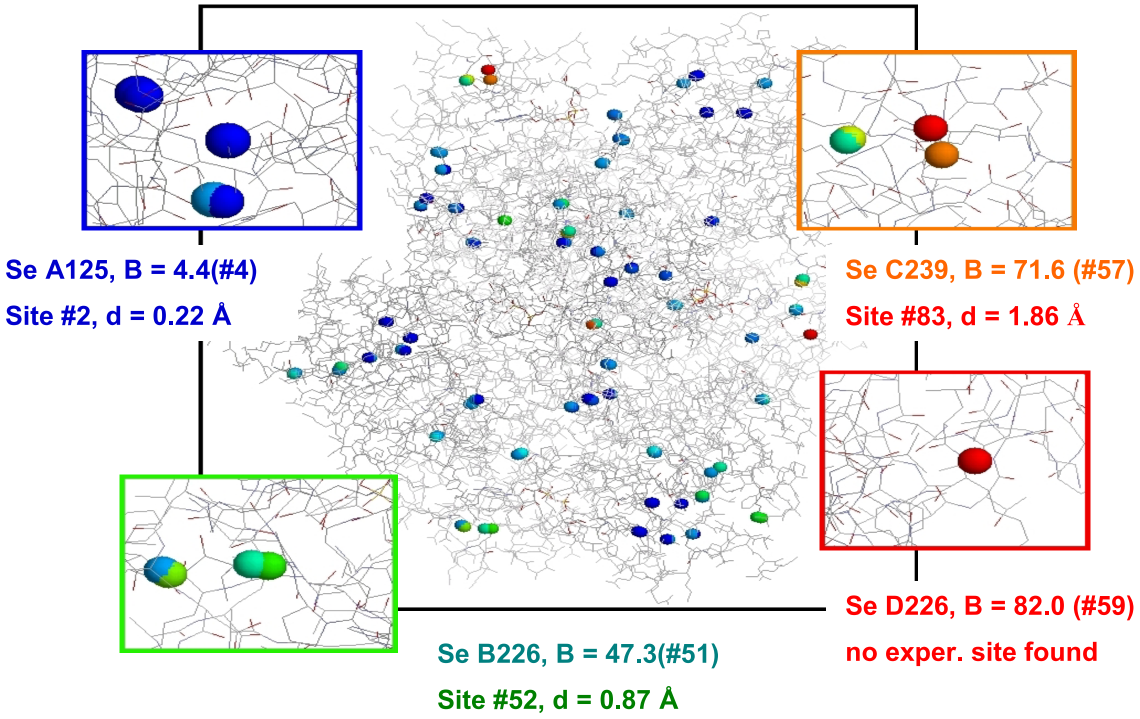


Fig. 4.48: The locations of the refinement/F_A-trial pairs in the THB structure. The colours correspond to B-values of the Se atoms and inverted peak heights of the sites (scaled to fit the B-value range), respectively. Examples for positionally well, less well and badly defined pairs are given with explicit values.

Also looking individually at the refined atom positions (Fig 4.48) a general agreement between the selenium B-values and the distance to the trial site partners can be found: In none of the cases, all 59 refined positions can be found by SHELXD, which is due to the fact that Se #59 (D226) has an exceptionally high B-value of 82.0 Å^2. The lower the displacement of the atoms, the higher is the corresponding experimental site peak height and the closer are refined and experimental position.

Large positional deviations in the high-B-value region can be found for all data subsets, but in the case of ΔF_{ip}, there are large disagreements already for refined atom #6 (not found at all) and #10, for which the weak site ip#57 is the partner at a distance of 1.9 Å.

Like Fig. 4.46 and Table 4.10, also Fig. 4.47 shows that for a SHELXD substructure solution at a classical threshold of 3.0 Å, the order of wavelength-dependent substructure quality is $F_A > \Delta F_{hr} > \Delta F_{peak} > \Delta F_{infl}$.

4.2.3 SAD phasing of HAPTBr

4.2.3.1 Introduction

HAPTBr is the structure of human *Acyl-Protein Thioesterase I* (Devedjiev et *al*. 2000), as determined from *SAD* phasing on a bromine-soaked crystal. High-energy remote *ΔF* data with a resolution limit of 1.8 Å was used to solve the structure. A native data set to 1.5 Å had been collected as well. The programs SNB for heavy atom structure solution (7 bromine sites found), SHARP for heavy atom model refinement (16 additional sites found from residual maps), SOLOMON for solvent flattening and ARP/WARP for the protein backbone autotracing were used. The structure determination had been completed using Refmac for model refinement, after which a total number of 40 bromine sites with varying occupancies were found.

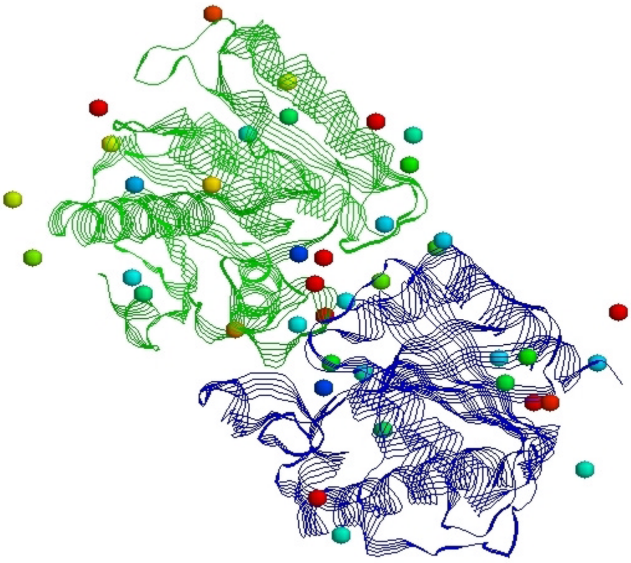


Fig. 4.49: The tertiary structure of hAPT-Br displayed in strand style. bromine sites are shown as spheres with a temperature colour scheme. HAPTBr crystallizes as a dimer and consists of 229 amino acid residues per monomer.

HAPTBr was chosen to test SITCOM in a realistic situation of a novel structure determination process. Therefore, in some of the present studies, it was pretended that *HATPTBr* was an unknown structure. The main difference to all “technical” test structures (Table 4.9) is the fact, that *HAPTBr* is not a selenium derivative with a (in principle) known number of sites, but a bromine soak structure, where neither the number of sites nor their occupancies are known a-priori. In fact, even the refinement result of 40 bromine atoms is not necessarily 100 per cent true. All studies on *HAPTBr* were done with *SAD* data from the peak wavelength subset, kindly provided by *Zbigniew Dauter*.

4.2.3.2 Overview

The following studies are divided in two major parts. First, the refined bromine positions are validated by test phase calculations with SHELXE, based on different selections of refined sites and a variation of (a) ΔF data and (b) ΔF *and* native *F* data resolution limits.

In the second part, experimental sites are re-determined with SHELXD. The substructure solutions from different ΔF data resolution limits are validated by (a) the comparison to the refined atom positions using SITCOM and (b) test phase calculations with SHELXE using full data resolution. The effect of solution selection and site filtering on phase quality is investigated, thus also judging the potential of the SITCOM solution cross-comparison method.

In both cases, SHELXE electron density maps were used for auto-tracing with ARP/WARP (Perrakis et *al*. 1999), focussing on the correlation between the phase errors and the interpretability of the experimental electron density map by the program.

4.2.3.3 The HAPTBr structure solution protocol

SHELXD was used in standard operation mode (as explained in the methods section, chapter 3.1.4), using the *PATS* instruction for patterson seeding of atoms. The resulting 100 solution trials, each consisting of 56 bromine sites, were analyzed with SitCom to select an optimal solution with a limited number of reliable sites. These sites were then supplied to the program SHELXE, run with 25 cycles of density modification (*-m25*) and an empirically optimized solvent content of 45%, (the true solvent content is 30%). The structure solution process was completed by protein autotracing, using a standard ARP/WARP (“warp’n’trace” mode) setup with 50 building cycles, each consisting of 5 Refmac sub-cycles.

Lacking the original *HAPTBr* native data, a pseudo-native data set (needed for SHELXE phasing) with Friedel-merged reflections was created from the peak data subset using the program XPREP. Of course, these data had the same resolution limit of 1.8 Å, and not the 1.5 Å of the true native data.

To evaluate the phasing results based on the different bromine substructures, reference phases were taken from a refined structure. For this purpose the original PDB model was refined against the previously created pseudo-native data using the program SHELXL (default CGLS refinement applying the *STIR* option to slowly include data and allow a stable model adaption to the SHELXL treatment). Phase comparisons were done, as usual, with SHELXPRO (see methods section).

4.2.3.4 *Test calculations using refined sites*

To estimate the principal effect of substructure completeness on phase quality (at different anomalous data resolutions), a decreasing number of refined bromine atom positions and occupancies were transformed from the PDB format to “pseudo” SHELXD output files, which served as input to SHELXE. SHELXD deduces occupancies of found heavy atom sites from their peak heights. Unlike SHARP, it does not refine B-values. Exploiting the fact that occupancies and B-values are correlated, it was tried to include the atom displacement information into the occupancy values of the pseudo-SHELXD files. It seemed that this information would be particularly important because of the fact that the given occupancies had been set to different values, but had not been refined.

The B-value correction was achieved by scaling the occupancy values to fit a fixed u value of 0.2. For example, the bromine atom #17, having a fixed occupany of 0.5 and a B-value of 23.3 (u = 0.295), becomes a SHELXD site with an occupancy of 0.3388. Like in a real SHELXD file, the pseudo sites were sorted by their occupancy values. It turned out, that the 40 sites with B-value corrected occupancies lead to much better phasing results than the same 40 sites with the original PDB occupancy values. This side result emphasizes the importance of free occupancy refinement, both for heavy atom substructures and for crystallographic structures in general.

Three sets of sites were finally tested as SHELXE input: Those taken from all 40 refined positions, a smaller set from the 20 strongest sites and a third one with the 10 strongest sites only. Additionally, the effect of anomalous data resolution on SHELXE results was studied. Therefore, the ΔF data were first used without resolution cutoff and then truncated to 2.0, 2.1, 2.2 and 2.5 Å.

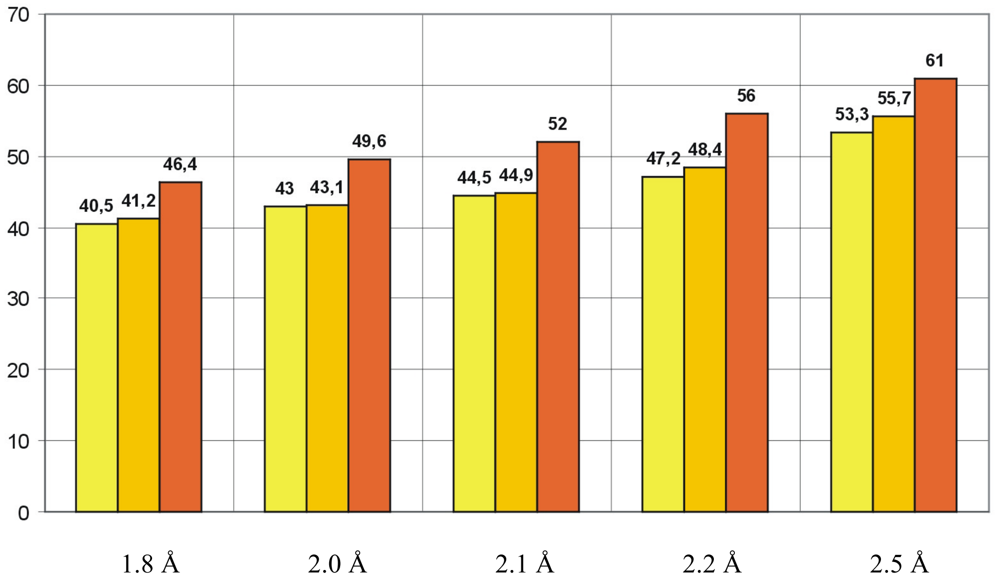


Fig. 4.50: SHELXPRO (F*FOM-weighted) phase error results for SHELXE phasing/density modification jobs (-m50) using different anomalous data resolution limits and different numbers of refined bromine substructure atoms: 40 sites (yellow), 20 sites (orange) and 10 sites (red-orange).

Two conclusions can be drawn from the results (Fig. 4.50): Apparently, there are no significant differences between the 20 and 40 site substructures, whereas the 10 site substructure is considerably worse with respect to phase error. This phenomenon is independent from resolution, meaning that only half of the sites are strong enough (or refined precisely enough, respectively) to contain a significant phasing information. It is rather irrelevant, how many of the the weaker sites are added to the critical number of twenty. Of course, this statement has to be done with a minimum amount of scepticism, because it should not be forgotten that the refinement is not expressing the perfect truth. It also has to be kept in mind that the reference phases of these studies are derived from the same refined model as the sites, so that a certain bias effect can be assumed for these results. Anyway, the result agrees quite well with the fact, that the number of 23 sites found by SHARP during the original structure solution (see above) has been of the same magnitude.

The second important fact learned from the diagram is the existence of a clear SHELXE dependence on anomalous data resolution: The resulting SHELXE phase errors increase almost linearily with the excess of resolution truncation. The need to understand the SHELXE dependence on data features gave rise to a more detailed study on data resolution.

4.2.3.5 Variation of resolution limits for native and anomalous data

This study, done with a substructure of 56 unmodified SHELXD sites, confirms the trend of results obtained before. Although a newer SHELXE version was used (*-m 100*), the increase of phase errors due to anomalous data truncation (Fig. 4.51, left group of bars) is comparable and only a little less steep than in the previous figure. The value of Δ(Δphi) is 4.0° when comparing 1.8 and 2.2 Å data limits, where it has been about 7° before (both for the 40 and 20 site refinement substructures). Truncating both anomalous and native data, the performance of SHELXE suffers significantly stronger and the increase of phase error is 10.8° in total.

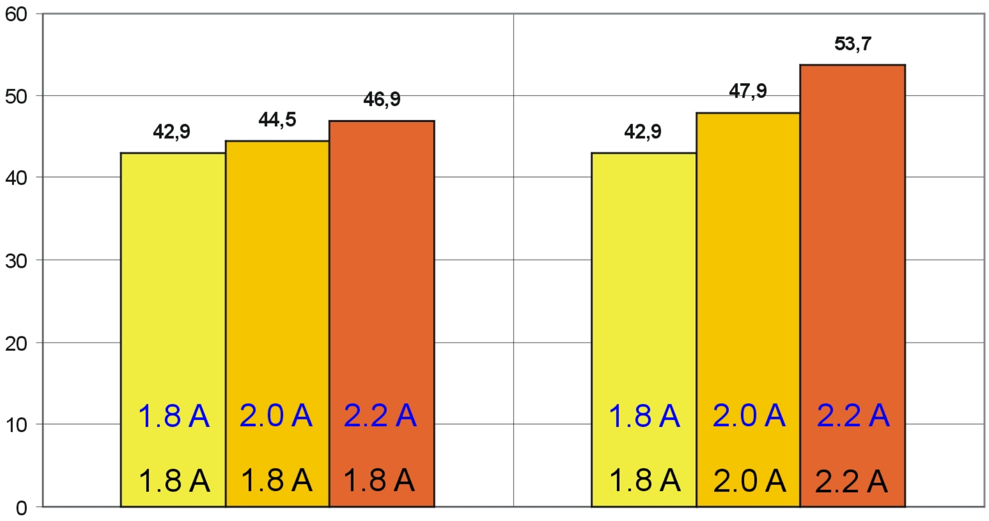


Fig. 4.51: The effect of anomalous and native data truncation (blue and black resolution limit values, respectively) on SHELXE phase quality, as determined with SHELXPRO. Yellow bars: no data truncation Orange bars: data truncation (anomalous or both) to 2.0 Å. Red-orange bars: data truncation (anomalous or both) to 2.2 Å. The substructure used was taken from the best solutution of a standard SHELXD with 2.0 Å anomalous data (56 sites). SHELXE was run with 100 cycles of density modification.

4.2.3.6 Auto-tracing results obtained from refined sites

Experimental phase errors are important hints at the resulting map quality and thus at the probabilty to solve the whole protein structure. From the practical point of view, the ability of a suited program to trace a sufficiently complete protein fold in the supplied map, is a more concrete and valuable phase quality indicator. Therefore, the application of the tracing program Arp/wArp completed each structure solution attempt, focussing on the correlation between SHEPLXPRO phase errors and number of traced residues after 50 Arp/wArp auto-building cycles. A general study (Fig. 4.52) was done using the same combinations of input substructures (40 or 20 sites from the refinement) and anomalous data resolution limits as before. ARP/WARP was run in “warp’n’trace” mode with default settings, using the “limited-Depth first” algorithm for α-carbon recognition and 5 cycles of the refinement program Refmac per building cycle. The experimental data supplied to ARP/WARP had been expanded to full resolution by SHELXE, because the native data used had been untruncated (1.8 Å).

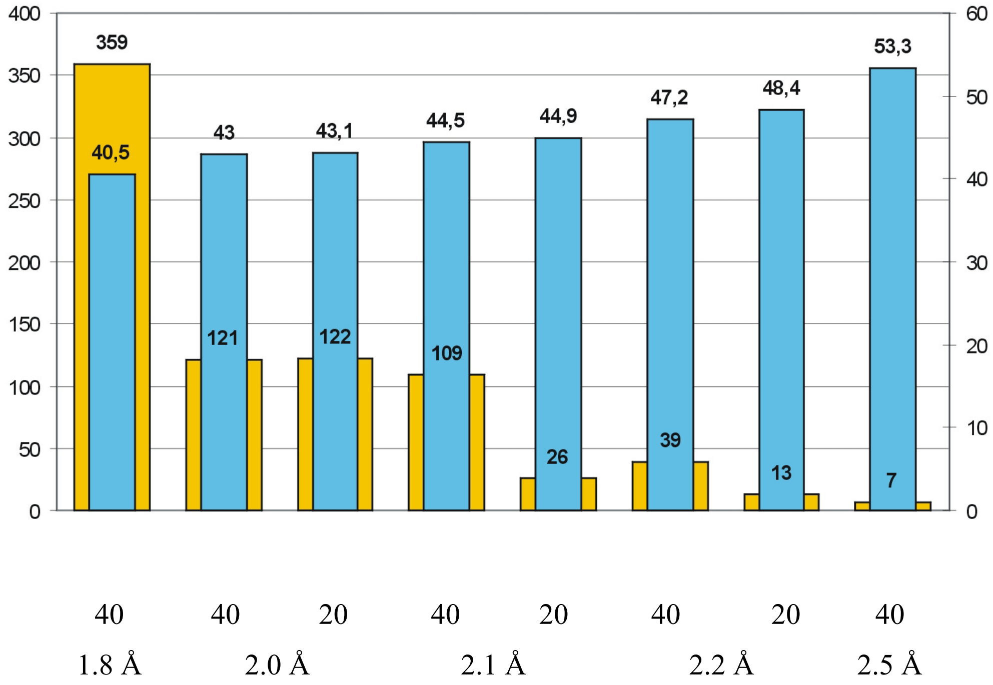


Fig. 4.52 SHELXPRO phase errors (blue bars) and numbers of residues successfully traced with Arp/wARP (orange bars) for SHELXE jobs run with 40 or 20 refined input sites at different limiting anomalous data resolutions. The complete HAPTBr model contains 458 amino acid residues per asymmetric unit.

Reducing the anomalous data resolution as major factor and the substructure completeness as minor factor (as discussed for Fig. 4.50), the phase errors increase slowly with a continuous, almost linear increment. As one would expect, the trend for the numbers of protein residues traced from the resulting experimental electron density maps is antiproportional to the phase error. There are, however sudden steps in residue numbers when getting from 40.5° to 43° and from 44.5° to 44.9° phase error. Only the complete substructure and the use of fully resolved anomalous data in SHELXE leads to a almost (78%) complete protein model after tracing. The numbers of residues obtained from the maps with 43 – 44.5° phase error correspond to 25% model completeness only, but experience has shown that a number of 40 – 50 residues after the first Arp/wArp attempt is often sufficient to expand the trace later, applying subsequent model building jobs with more than 100 cycles. The Arp/wArp results for maps with 45° phase error and worse are below a number of 40 residues, however, and do not seem likely to lead to any protein structure solution.

The most striking – yet open – question is, why the seemingly insignificant phase error difference of 0.4° between the 40 and 20 input site SHELXE jobs with 2.1 Å resolution limit leads to the Arp/wArp difference between a rather successful 109 residue trace and a hopeless 26 residue trace. More detailed analyses of both the data characteristics (phase error, FOM and map CC profiles along the resolution range) and the Arp/Warp building protocols are needed to answer this question in the future.

4.2.3.7 Analysis of SHELXD substructure accuracy

While SHELXE depends quite critically on anomalous data resolution, SHELXD is known to be very robust for a wide range of resolution limits. Still, it was tried to analize the effects of using different *SHEL* values, truncating SHELXD data to 2.0, 2.5, 3.0 and 3.5 A. After each job, SHELXD selects the solution with the highest correlation coefficient, recommending those sites for use with SHELXE. Assuming 40 sites as estimated heavy atom number (*FIND 40*), all SHELXD jobs produce 56 possible sites.

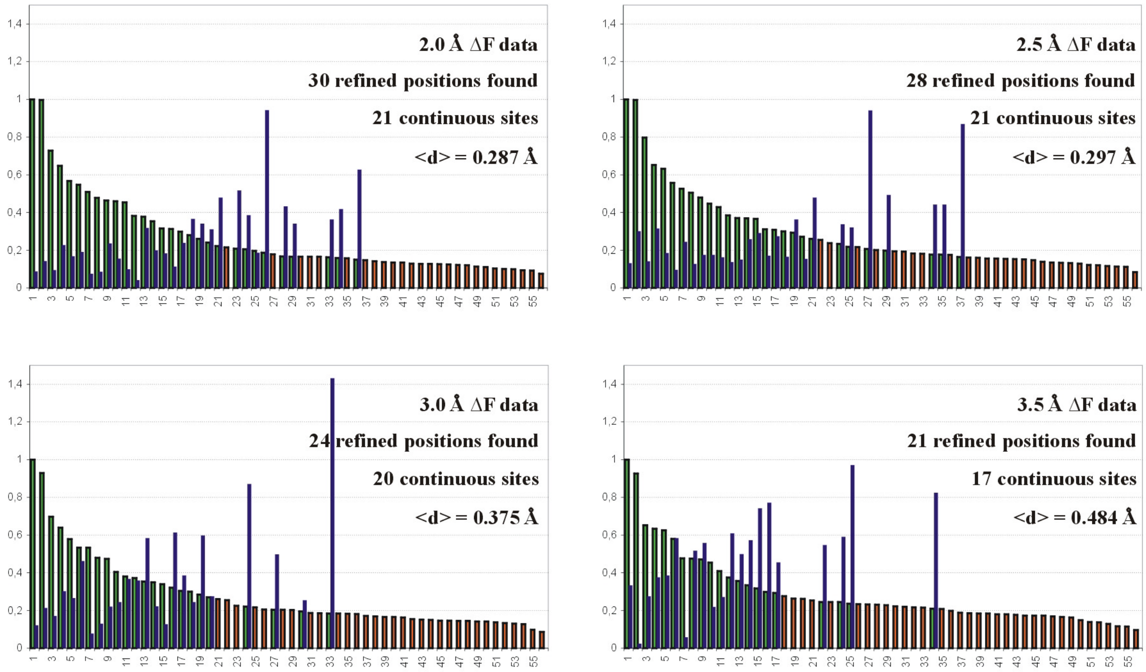


Fig. 4.53: The four SHELXD substructures with highest correlation coefficients, resulting from jobs at 2.0, 2.5, 3.0 and 3.5 Å. SHELXD site occupancies (green and orange bars) and distances to refined Br atom positions (blue) are plotted against site number. Sites without corresponding refinement position are coloured orange.

Looking at the four SHELXD subtructures from the solutions with highest correlation coefficients, a virtually identical trend (decrease) of occupancies is found, independent from the wavelength limits. After the second site of each substructure, a clear step of about –20% is observed, but apart from this, the occupancy falloff is rather flat and continuous. Judging from the occupancies alone, one could hardly decide how many sites to keep for phasing. Interestingly, the substructures also lack a clear step around site 20, the apparent true number of strong sites (see previous studies), or around site 40, the number including weak sites according to refinement. Thus, for the subsequent phasing validation, all 56 sites were supplied to SHELXE, as one would of course do without prior knowledge of the soaked bromine atom number.

The comparison to refined bromine atom positions reveals a loss of substructure accuracy with lower resolution limits. The best solutions at 2.0 Å and 2.5 Å are still very similar. For the 3.0 Å solution and to a larger extent for the 3.5 Å solution, all accuracy indicators become significantly worse – the total number of found "true" positions as well as the number of subsequent (continuous) refinement-consistent sites and the mean distance of sites to the reference positions.

4.2.3.8 Phasing with SHELXD sites

Simulating the application of SITCOM in a novel structure determination, the sets of 100 SHELXD solutions per wavelength were evaluated in cross-comparison mode without the use of *a-posteriori* refinement information (see methods section, chapter 3.2). As the correlation coefficients were very similar, a high initial FOM selection limit of 95% was used, e.g. keeping only the best three SHELXD solutions in case of the 3.0 Å job. For each job, the solution with the highest agreement to most other solutions *i.e.* the highest SitCom score was selected. From these best solutions, only sites *common to all analyzed sitelists* of the respective job were taken into the final SITCOM substructure and supplied to SHELXE.

For every substructure validated here, SHELXE was used with full resolution for both anomalous and native data and with 25 cycles of density modification.

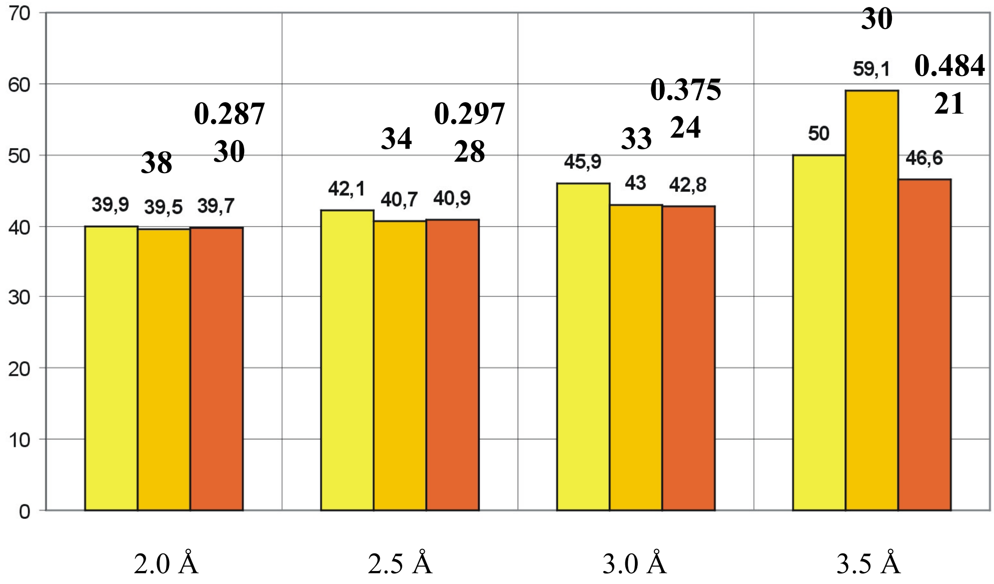


Fig. 4.54: Phase errors for standard (1.8 Å, 25 DM cycle) SHELXE jobs supplied with **(a)** 56 SHELXD sites from CCmax solutions (yellow bars), **(b)** SitCom-filtered SHELXD sites (present in all tested solutions) from the cross-comparison substructures with highest scores (orange bars) and **(c)** SHELXD sites agreeing best to the refined bromine atom positions after SitCom reference comparison (red-orange bars). The bars are grouped by the SHELXD resolution limit used. The numbers remaining of sites per substructure are given for the all SITCOM-treated cases. For the substructures compared to the refined positions, the mean distances [Å] are given as well (this information has also been given in Fig. 4.53).

Looking at the SHELXD resolution effect on substructure phasing accuracy, a steady increase of phase error values for all three sets (types) of SHELXE input sites can be observed when reducing the resolution. This effect is however rather small up to a limit of 3.0 Å in SHELXD.

The overall resolution-dependent phase error increase reflects the substructure accuracy development investigated previously (Fig. 4.54). A even more interesting aspect is the phase difference between substructures treated with or without SITCOM. There is virtually no phase error difference for the three site types at 2.0 Å, and only a very small improvement of roughly 1° at 2.5 Å, but with a 3.0 Å resolution limit, where SHELXD substructure accuracy starts to suffer, a decrease of approximately 3° phase error can be achieved using SITCOM in either refinement comparison or – more important – in "knowledge-free" cross-comparison mode. Taken alone, the 3° phase error improvement may seem small, but as we already know, small phase changes can have astonishing effects on the traceability in the resulting experimental map.

4.2.3.9 Autotracing results obtained from SHELXD sites

There is a remarkable benefit in Arp/wArp protein autobuilding performance after treatment of the relatively best SHELXD substructure solutions with SITCOM. Looking at Fig. 4.55, the maps from the initial 2.0 Å SHELXD substructures, having phase errors slightly below 40°, are by far the best ones for autotracing, leading to protein models with more than 80% residue completeness. There are only small relative differences between the types of substructures. Still, the one from SITCOM cross-comparison is the best of the tree, like it is the best also at 2.5 and 3.0 Å. For these two SHELXD resolution limits, the two respective SITCOM substructures cause considerably better autotracing results than the unmodified 56 site SHELXD solution. In the border case of 3.0 Å, the application of SitCom makes the difference between a completely failed Arp/wArp job with no residues built, and a successful 87 residue trace(19% complete), from which the building of the remaining fold should be possible. The refinement-compared SITCOM sites lead to a less successful, but still quite acceptable 56 residue trace. With a 3.5 Å SHELXD resolution limit, all phase errors are greater than 45° and all corresponding maps are hardly traceable. In this case, the resolution is obviously too low to get sufficiently precise substructures. In particular, the inconsistency of the SHELXD solutions causes the solution cross-comparison mode of SITCOM to fail – the phase error is highest and no residue can be traced in the experimental map.

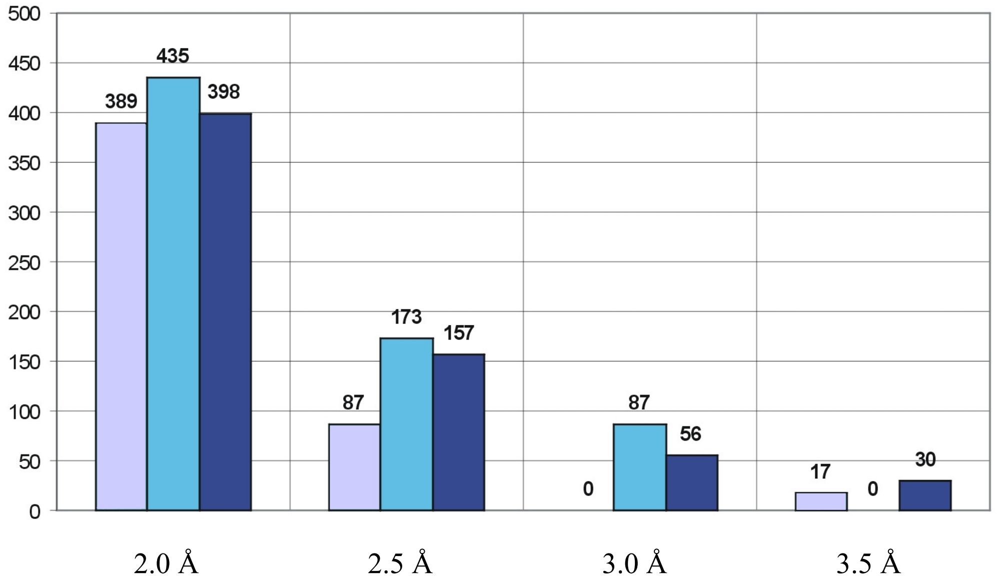


Fig. 4.55: Numbers of HAPTBr residues traced with Arp/wArp using the SHELXE maps characterized by phase errors in the previous figure. The order of bars is related to the substructure types the same way as before. Again, the resolution values refer to the limits used initially in SHELXD.

In case of *HAPTBr*, all SHELXD solutions of a given job had very similar *CC* values and were relatively site-consistent (except for the job limited to 3.5 Å resolution). Therefore, it can be assumed that the SITCOM selection of a solution scored slightly higher than the rest would not have made a big difference to another solution. For substructure accuracy improvement, the crucial operation is rather the selection of sites, filtering out those unreliable ones that are found in few solutions only. As we know from the preparative studies, about 20 bromine sites are strong enough too contain the relevant *HAPTBr* phasing information. The remaining sites do not only contribute less and less to phasing – they are also adding an increasing amount of systematic errors to the phase determination, as they can not precisely be determined. At least for the SHELXD sites beyond number forty, the introduction of noise can be expected to dominate any further structural information and thus to deteriorate the phase and experimental map quality. (Of course it can not be excluded that there might be more than 40 bromine atoms in the *HAPTBr* structure, although too weak to be detected).

From another point of view, there is a certain uncertainity also in the weaker refinement positions and occupancies of the finally determined bromine atoms. This explains, why the numbers of experimental sites corresponding to the refinement (due to the SITCOM reference comparison mode) are smaller than the remaining cross-comparison site numbers, and still lead to phasing and tracing results worse than without using refinement information.

It can be concluded that the value of using SITCOM after the initial substructure solution is connected to the site selection of a given solution rather than to the solution selection itself – this is at least true if a strict FOM value pre-selection is done in case of a largely differing CC distribution. Because of the dominant site-filtering role of the program, it can be predicted than the major field of SITCOM application will be heavy atom-soaked structures like HAPTBr instead of selenium derivatives which are in general solved by SHELXD / SHELXE (or similar programs) alone and without difficulties.

4.2.3.10 The Comparison of phasing / phase improvement programs

Completing the studies on the HAPTBr structure solution, the map quality differences between phasing programs on the one hand, and the connection between map quality and autotracing success on the other hand shall be highlighted.

In a pure SHELX structure solution process, SHELXD substructures are supplied to SHELXE without further positional or other parameter refinement. Alternatively, the coordinates, occupancies and ADPs of a given set of SHELXD sites can be refined by SHARP and then used by this program to calculate preliminary protein phases. The density modification part of SHELXE treatment is done by SOLOMON. The SHARP/SOLOMON structure solution was carried out with some variations in order to estimate the importance of substructure model refinement as well as the correctness of the fixed occupancies determined with SHELXD. In this case the 2.0 Å SHELXD sites, treated with SitCom in cross-comparison mode, were taken as initial substructure source for all phasing procedures. Both native and anomalous data were used with full 1.8 Å resolution, in case of SHARP as a combined MTZ file. Both SHELXE and SOLOMON were run with 50 cycles of density modification and a solvent content of 45%.

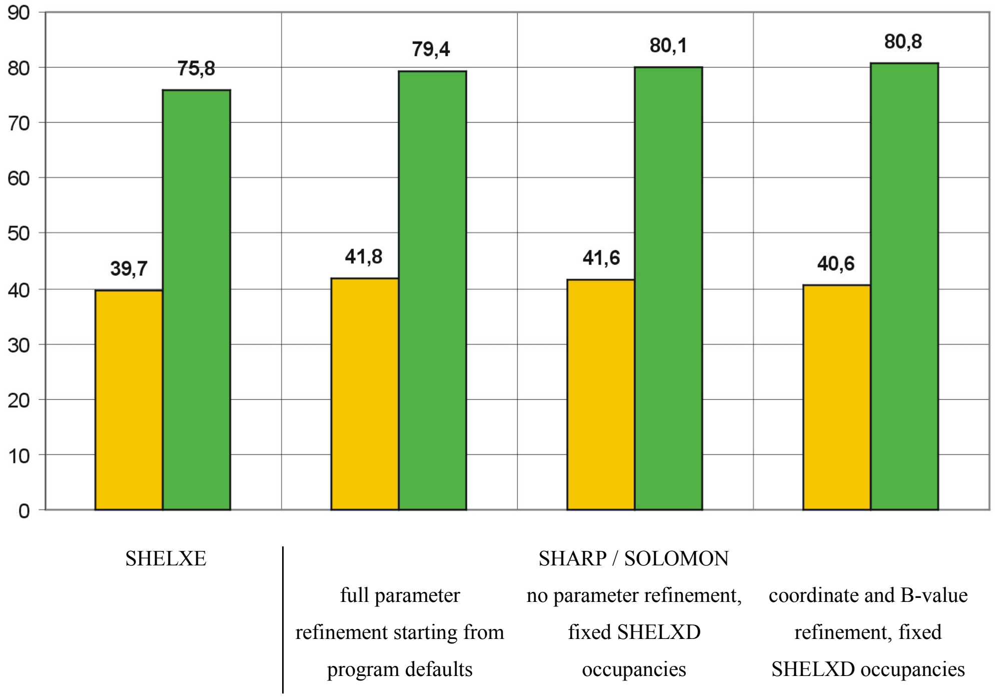


Fig. 4.56: Phase error (yellow) and map correlation coefficient (green) values for experimental maps obtained from SHELXE and SHARP/SOLOMON under different working conditions.

Looking at the phase errors, surprisingly all SHARP / SOLOMON results are worse than the SHELXE result, although not dramatically. The reciprocal space map correlation coefficients, however, are about 4 – 5% better. Between the variants of SHARP operation, there are no significant phase quality differences, but judging from the small improvement of values, it is better to use fixed SHELXD site occupancies than to refine them from a default start value of 30. Combining these fixed B-values with a refinement of the remaining parameters, the best phase error and CC values are obtained.

In general, the similarity of SHARP results leads to the conclusion that the main contribution to map CC improvement relative to SHELXE does not lie in heavy atom model refinement but rather in the phase calculation method of SHARP or in the density modification method of Solomon. To distinguish these two factors, comparative phasing sessions should be done in the future, using SHELXE without density modification and SHARP without SOLOMON, respectively. Looking at the shape of the experimental maps, the SOLOMON map connectivity is better than the SHELXE connectivity.

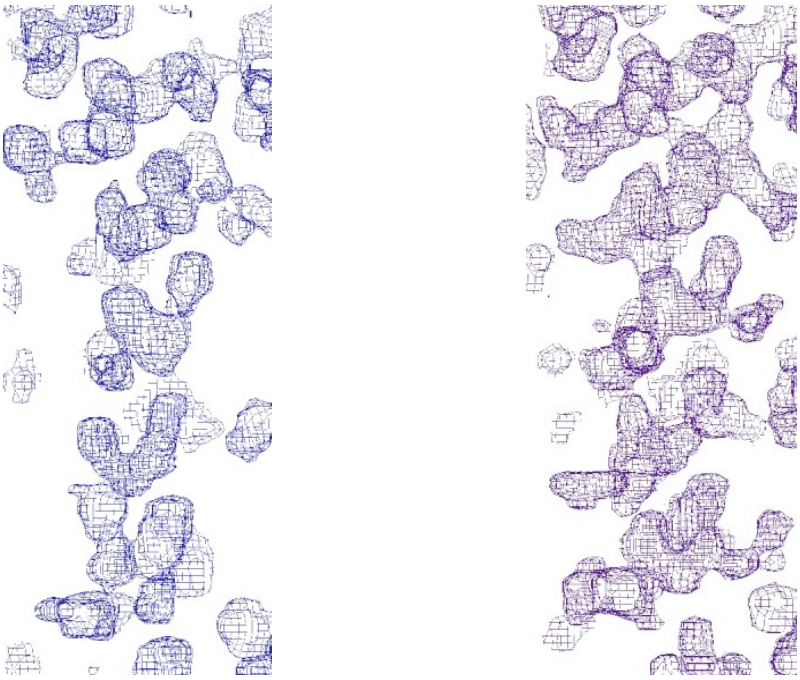

(a) **(b)**

Fig. 4.57: Experimental (F_o, phi_{MAD}) electron density, contoured at a level of 1 σ, in the region of a HAPTBr α-helix. **(a)**: the SHELXE map. **(b)**: the SHARP/SOLOMON map

The ARP/WARP autotracing result obtained using the SHARP/SOLOMON map is the best of all the HAPTBr studies, not only because of a final value of 442 traced residues, but also because of the fact that virtually all residues are found within the first 10 cycles (see graph in Fig. 4.58). It is not sure, whether the higher map CC value or the higher connectivity of the electron density causes the improved (automated) interpretability of the SHARP/SOLOMON map. As the ARP/WARP tracing algorithm is based on Cα recognition, a less connected electron density with higher "atomicity" should have no negative influence on the autobuilding success.

map source	Δ(phi)	map CC	Arp/wArp residues (10 cyles)	Arp/wArp residues (50 cycles)
SHELXD (3.5 Å) / -E (1.8 Å)	50.0°	0.620	17 (4%)	17 (4%)
40 refined sites / SHELXE (2.2 Å)	47.2°	0.686	27 (6%)	39 (9%)
SHELXD (2.5 Å) / -E (1.8 Å)	42.1°	0.727	73 (16%)	87 (19%)
SHELXD (2.0 Å) + SITCOM SHELXE (1.8 Å)	39.5°	0.759	224 (49%)	435 (95%)
SHELXD (2.0 Å) + SITCOM SHARP / SOLOMON (1.8 Å)	40.6°	0.808	426 (93%)	442 (97%)

Table 4.11: some combinations of site sources and phasing conditions from the previously presented HAPTBr studies, leading to different maps. The map validation by the SHELXPRO quality indicators is compared to the Arp/wArp results.

Besides the already drawn conclusions about SitCom, one can learn from the *HAPTBr* studies that all data available (i.e. full resolution) should be supplied both to SHELXD and SHELXE, as long as the data quality, like for the HAPTBr peak data, has a normal resolution dependency. SHELXE depends stronger on anomalous high resolution data than SHELXD, but also small differences in substructure accuracy can have large effects on the final tracing results. The use of SitCom is always recommended for heavy atom soak structures.

The auto-tracing behaviour of ARP/WARP can not be reliably predicted in every case, but the map correlation coefficients, more than the phase errors, are a good hint.

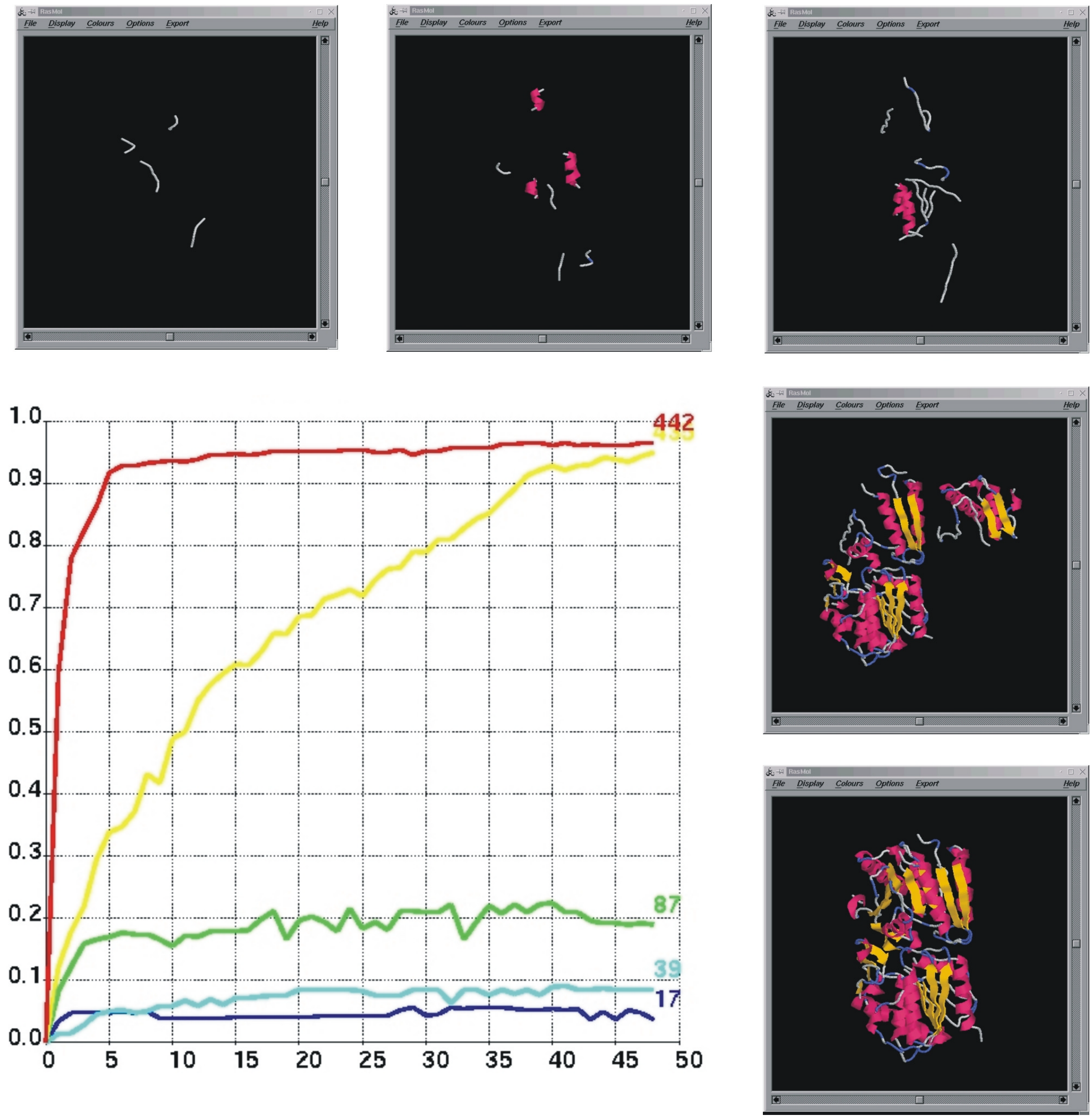


Fig. 4.58: The development of the ARP/WARP tracing models for the five exemplary map sources presented in Table 4.11. Model completeness is plotted against autobuilding cycles. The final residue numbers are given explicitly, and the final models are displayed in the Rasmol images, sorted from top left to bottom right.

5 Summary and Conclusions

5.1 Aldose Reductase

The structure of human Aldose Reductase in complex with the inhibitor IDD594 was used as a model system for high-resolution phasing. From the three-wavelength *MAD* data resolved to 0.9 Å, phases of extraordinarily high quality were obtained after the selenium substructure determination and the refinement of the heavy atom parameters. This was indicated by a very small deviation between the experimental phases and the model-calculated phases (less than 20° phase error). Thus, it was shown that the very high accuracy of signed anomalous differences for all three data subsets, exhibiting correlation coefficients of about 80% overall, and well above 30% even for the outer resolution shell, allows to derive experimentally phased electron density maps in a straight-forward fashion without the need for any intermediate phase improvement by means of density modification methods.

The refinement of the *hAR2* model against structure factor intensities from the high-energy remote data yielded final *R* values of 7.6% (R_{work}) and 9.0% (R_{free}), indicating a very good agreement between the refined model and the observations. From the opposite point of view, the unproblematic progress and the very good results of the refinement also reflect the high data quality in terms of merged Friedel intensities, at least for the data collected at high-energy remote wavelength, which showed a particularly high completeness and intensity-over-sigma. The refinement against data resolved to 0.9 Å allowed the identification of conformational disorder for 67 amino acids, including networks of connected sidechains and backbone shifts, as well as the detailed characterization of positionally fixed solvent water molecules, for many of which partial occupancies were observed (and in some cases refined). The geometry of the active site found by previous structure determinations was confimed, especially with respect to the hydrogen bond interactions relevant for ligand and coenzyme binding. The hydrogen atom recognition using F_o-F_c difference density maps, in particular at residues involved in the active site, was unsatisfying, because significant density peaks were missing at many expected polar hydrogen atom positions.

The experimentally phased electron density map was applied to validate the refined model. A systematic analysis and classification of disordered amino acid residues was made, thus implying also the assessment of the model-calculated electron density maps used during the refinement, which could not be regarded as bias-free. Most of the modelled disorder positions as well as the assignments of primary and secondary conformations with respect to the refined occupancies were confirmed. This leads to the conclusion that the conventional strategy for disorder refinement at high resolution, using difference density maps from calculated phases, is generally reasonable and leads to correct results that agree with the *purely* experimental observations. Very flexible regions of the protein, such as the C-terminus or amino acid side-chains pointing into the bulk solvent area were hardly traceable also in the experimental electron density. The attempt to localize hydrogen atoms in the experimental map has not been satisfying. The question whether this is merely the result of still lacking phase quality at the very high resolution edge or a more fundamental problem in protein crystallography is open to discussion.

The experimentally phased map from *MAD* data has proven to be suited for the purpose of model verification at atomic resolution. This result agrees with other studies recently made for different, but comparable structures (e.g. Schmidt et *al.* 2002) and marks the method as a way to overcome the model bias problem in protein crystallography. The potential of atomically resolved experimental maps for complete structure determination (i.e. without refinement) has not been studied, but it is well justified to assume that a map suited for the *a posteriori* evaluation of model details would be very beneficial for "*ab initio* modeling" as well.

5.2 SitCom

The program SitCom for heavy atom substructure comparison and validation was developed and applied to several test structures, as well as to *human Acyl Protein Thioesterase* serving as a model structure for studies on halide-soak based *SAD* phasing. The program proved to be fully functional in all principal types of spacegroups. Furthermore it was shown that SitCom can be profitably used both for the *a posteriori* substructure assessment and for novel structure determination with heavy atom derivatives.

5.2.1 Studies on Transhydrogenase B

For the *Transhydrogenase B* (*THB*) structure, heavy atom substructure solutions obtained from SHELXD jobs against F_A data and three sets of ΔF data from the three *MAD* wavelengths, were compared to the refined selenium positions. The comparisons were done for different data resolution limits. The studies revealed a comparably high accuracy of the substructures determined from F_A data and ΔF data from peak and high-energy remote. For all of these data types, a sufficient percentage of correct solutions were found, which contained sites corresponding to most of the refined selenium positions, with convincingly close distances well below 0.5 Å.

Especially for the single wavelength data subsets, the results underlined the importance of the correct choice of individual data resolution cutoffs, as it proved to be beneficial for the substructure accuracy to use a high resolution limit of 2.5 Å, i.e. many data of the peak ΔF subset showing a high $I/\sigma(I)$, but disadvantageous to keep this limit for the other two subsets, which exhibited lower $I/s(I)$ and therefore better substructure results with lower resolution cutoffs.

The relative significance of the *E*-value-based correlation coefficients $CC(E_{obs}, E_{calc})$ was confirmed. SHELXD solution distributions are usually bimodal in terms of correlation between *CC* values and Patterson-based figures of merit (*PatFOM*) for the heavy atom sites. The results for *THB* showed that the substructures corresponding to the group of solutions with highest *CC* values were most likely to be correct. However, the *CC* values strongly depend on the quality of the structure factors involved. They are higher for F_A data, being more accurate structure factor estimates, than for the ΔF data. The *CC* values are as well higher for stronger data with lower resolution cutoffs than for weaker data with higher resolution limits. Therefore, their absolute magnitude cannot be taken as indicator for the substructure correctness. In the *THB* case, some solutions from F_A data at 3.5 Å with *CC* values greater than 50% were found to have no site at all within a 2 Å radius to refined atom positions, while other solutions from high-energy remote data at 2.0 Å with CC values of less than 30% were correct, representing relatively complete and accurate substructures.

5.2.2 Studies on human Acyl Protein Thioesterase I

HAPTBr is a structure containing bromine atoms from a soak of the crystal immediately before the experiment, therefore the number of the sites is not known and their occupancies vary. The accuracy of unmodified SHELXD substructures was assessed by *a-posteriori* comparison to refined bromine atom positions (as done for *THB*). Additionally, the quality of the unmodified substructures as well as of substructures treated with SITCOM was validated by the effect on SHELXE phase reliability and subsequent model building success.

In initial test studies, it had been shown that only the 20 refined bromine atoms with the highest occupancies (out of a total number of 40), when submitted to SHELXE, are responsible for relatively reliable phases, deviating about 40° from refined phases.

For the 56 sites determined by SHELXD, the occupancies have been found to decline very continuously, lacking a significant step that would indicate a border between correct and wrong sites. This observation was made for all resolution cut-offs. The *a posteriori* evaluation by comparison to refined atoms revealed that the substructure accuracy was highest for data truncated to 2.0 Å resolution, while it decreased upon lowering the cut-off values. This result, reflecting a high $I/\sigma(I)$ is similar to the one obtained for *THB* substructures from peak ΔF data. In general, the pairs of corresponding positions were related to the strongest 20-24 sites of both the refined and the experimental set.

Picking experimental sites consistent to all other (selected) solutions of the same SHELXD job, SITCOM discarded the 18 – 26 weakest ones (depending on the resolution cut-off used) of each best substructure. The remaining sites have proven to correspond to the strongest refined atoms as well, except for the 3.5 Å substructure which seemed to be systematically incorrect.

By comparison to refined phases it was shown that the phases obtained from standard SHELXE jobs provided with the selected experimental sites were about 3° more accurate than those resulting from the unmodified SHELXD substructures. This beneficial effect of site selection was even more pronounced in the auto-building results based on SHELXE electron density maps, where previously failed tracing attempts became successful. The phasing and tracing results support the assumption that the weaker SHELXD sites are wrong and therefore introduce noise into the protein phases.

Thus, the studies on *HAPTBr* highlight the need for site correction after SHELXD in heavy atom soak cases and the potential of SITCOM to improve the heavy atom substructures and the resulting phases by finding sites consistent to many solutions.

5.3 Future perspectives and final remarks

The need for an experimental proof of the hydrogen atom state in the active site of *hAR2* will require further efforts in the exploitatation of experimental phases at sub-atomic resolution. As the power of crystallography is generally limited in this respect, the use of alternative methods, in particular neutron diffraction, has to be considered (Engler et *al.* 2003) – however, this would require crystal sizes which can hardly be achieved from protein samples.

The availability of a purely experimental electron density map, free of any model assumptions, should in principle allow the analysis of the bulk solvent region which is normally considered as “flat”. The applicability of the present 0.9 Å MAD map for “ab initio modeling”, exceeding the level of details obtained from normal autotracing, has already been mentioned.

The positional comparison of the *hAR2* models obtained from the refinements against the three individual data subsets (high energy remote, peak and inflection point) at 1.5 Å, might serve as an indicator for the correlation between these subsets. Such a comparison could be done using distance matrices (Schneider 2002). The differences between the models might also reveal valuable information about radiation damage and help to answer the question of which data subset best to take for structure refinement.

The results obtained from the *a posteriori* substructure evaluations with SITCOM, especially for *THB*, have pointed out the possibility of using ΔF subsets from the single wavelengths of a MAD experiment for successful substructure determination. This supports the suggestion to attempt the structure solution already during the MAD experiment, after completing the collection of the first data subset, and to skip the experiment in case of premature success (Dauter 2002).

The application of SITCOM in the phasing process for heavy atom soak structures should facilitate the solution of difficult structures in the future. In this context, the program should also be tested on critical structures, which can normally not be solved by the usual substrucure determination programs.

The selection of consistent heavy atom sites has proven to be a very effective method for different solutions of the same program and the same job. Still, the risk of finding consistent sites of equally wrong solutions remains (see the 3.5 Å *HAPTBr* case). Therefore the comparison of presumably more independent solutions from different programs should be studied in more detail.

Finally, the analysis of non-crystallographic symmetry should be implemented in SITCOM, making the identification of correct sites more effective.

Concluding this thesis, it can be stated that the studies presented here have contributed to the improvement of methods for experimental macromolecular phasing and have additionally illustrated the benefits of experimental phasing at atomic resolution.

6 References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Methods used in the structure determination of bovine mitochondrial F1 ATPase*. Acta Cryst. **D52**, 30-42.

Agarwal, R. C. (1978). *A new least-squares refinement technique based on the fast Fourier transform algorithm*. Acta Cryst. **A34**, 791-809.

Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. (2000).*The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution*, Science **289**, 905 – 919.

Blow, D. M. & Crick, F. H. C. (1959). *The treatment of errors in the isomorphous replacement method*. Acta Cryst. **12**, 794-802.

Bohren, K. M., Grimshaw, C. E., Lai, C.-J., Harrison, D. H., Ringe, D. Petsko, G. A. & Gabbay, K. H. (1994). *Tyrosine-48 is the proton donor and histidine-110 directs substrate stereochemical selectivity in the reduction reaction of human aldose reductase: enzyme kinetics and crystal structure of the Y48H mutant enzyme*. Biochemistry **33**, 2021-2032.

Branden, C. I. & Jones, T. A. (1990). *Between objectivity and subjectivity*. Nature **343,** 687-689.

Bricogne, G. (1991). *A maximum-likelihood theory of heavy-atom parameter refinement in the isomorphous replacement method*. In *Isomorphous Replacement and Anomalous Scattering*., edited by W. Wolf, P. R. Evans, & A. G. W. Leslie, pp. 60-68. Proc. Daresbury Study Weekend, Daresbury Laboratory, Warrington, UK.

Brodersen, D. E., de La Fortelle, E., Vonrhein, C., Bricogne, G., Nyborg, J. & Kjeldgaard, M. (2000). *Applications of single-wavelength anomalous dispersion at high and atomic resolution*. Acta Cryst. **D56**, 431 – 441.

Brünger, A. T. (1992) *Free R value: A novel statistical quantity for assessing the accuracy of crystal structures*. Nature **355,** 472–475.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination*. Acta Cryst. **D54**, 905-921.

Buckley, P. A., Jackson, J. B., Schneider, T. R., White, S. A., Rice, D. W. & Baker, P. J. (2000). *Protein-protein recognition, hydride transfer and proton pumping in the transhydrogenase complex*. Structure **8**, 809-815.

Burling, F. T., Weis, W. I., Flaherty, K. M. & Brünger, A. T. (1996). *Direct observation of protein solvation and discrete disorder with experimental crystallographic phases*. Science **271**, 72-77.

Cachau, R., Howard, E., Barth, P., Mitschler, A., Chevrier, B., Lamour, V., Joachimiak, A., Sanishvili, R., Van Zandt, M., Sibley, E., Moras, D. & Podjarny, A. (2000). *Model of the catalytic mechanism of human aldose reductase based on quantum chemical calculations*. J. Phys. IV, 10.

Calderone, V., Chevrier, B., Van Zandt, M., Lamour, V., Howard, E., Poterszman, A., Barth, P., Mitschler, A., Lu, J., Dvornik, D. M., Klebe, G., Kraemer, O., Moorman, A. R., Moras, D. & Podjarny, A. (2000). *The Structure of Human Aldose Reductase bound to the Inhibitor IDD384*. Acta Cryst. **D56,** 536-540.

Collaborative Computational Project, Number 4 (1994). *The CCP4 Suite: Programs for Protein Crystallography*. Acta Cryst. **D50**, 760-763.

Cowtan, K. (1994). *'dm': An automated procedure for phase improvement by density modification*. Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography **31**, 34-38.

Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1997). *The benefits of atomic resolution*. Curr. Opin. Struct. Biol. **7**, 681-688.

Dauter, Z., Dauter, M., de La Fortelle, E., Bricogne, G. & Sheldrick, G. M. (1999). *Can anomalous signal of sulfur become a tool for solving protein crystal structures?*. J. Mol. Biol. **289**, 83-92.

Dauter, Z. & Dauter, M. (1999). *Anomalous Signal of Solvent Bromides used for Phasing of Lysozyme*. J. Mol. Biol. **289**, 93-101.

Dauter, Z. & Adamiak, D. A. (2001). *Anomalous signal of phosphorus used for phasing DNA oligomer: importance of data redundancy*. Acta Cryst. **D57**, 990-995.

Dauter, Z. (2002). *One-and-a-half wavelength approach*. Acta Cryst. **D58**, 1958-67.

Debreczeni, J. E., Bunkóczi, G., Ma, Q., Blaser, H. & Sheldrick, G. M. (2003). *In-house measurement of the sulfur anomalous signal and its use for phasing*. Acta Cryst. **D59**, 688-696.

van Delft, F., submitted to Acta Cryst.

Devedjiev, Y., Dauter, Z., Kuznetsov, S. R., Jones, T. L. Z. & Derewenda, Z. S. (2000). *Crystal Structure of the Human Acyl Protein Thioesterase I from a Single X-Ray Data Set to 1.5 Å*. Structure **8**, 1137-1146.

Drenth, J. (1994). *Principles of Protein X-ray Crystallography*. 2nd ed., Springer Verlag, New York, 1999.

Dvornik, D., Millen, J., Hicks, D. R. & Kraml, M. (1994). *Tolrestat pharmacokinetics in rat peripheral nerve*. J. Diabetes Complic. **8**, 18-26.

El-Kabbani, O., Ramsland, P., Darmanin, C., Chung, R. & Podjarny, A. (2003). *Structure of Human Aldose Reductase in Complex with Statil: An Approach to Structure-Based Inhibitor Design of the Enzyme.* Proteins: Structure, Function and Genetics **50**, 230-238.

Engler, N., Ostermann, A., Niimura, N. & Parak, F. G. (2003). *Hydrogen atoms in proteins: Positions and dynamics*. PNAS **100**, 10243-10248.

Esnouf, R. M. (1997). *An extensively modified version of MolScript that includes greatly enhanced coloring capabilities*. J. Mol. Graph. **15**, 132-134.

Fujinaga, M. & Read, R. J. (1987). *Experiences with a new translation-function program*. J. Appl. Cryst. **20**, 517 – 521.

Gonen, B. & Dvornik, D. (1995). *The sorbitol pathway and diabetic complications*. In *Diabetes - Clinical Science in Practice*, edited by D. E. Robbins & D. Leslie, Cambridge University Press, 313-330.

Hall, D. R., Gourley, D. G., Leonard G. A., Duke, E. M. H., Anderson, L. A., Boxer, D. H. & Hunter, W.N. (1999). *The high-resolution crystal structure of the molybdate-dependent transcriptional regulator (ModE) from Escherichia coli: a novel combination of domain folds*. The Embo Journal **18**, 1435-1446.

Harker, D. (1956). *The determination of the phases of the structure factors of non-centrosymmetric crystals by the method of double isomorphous replacement*. Acta Cryst. **9**, 1.

Hendrickson, W. A., Smith, J. L. & Sheriff, S. (1985). *Direct Phase Determination Based on Anomalous Scattering*. In *Methods in Enzymology*, edited by H. W. Wyckoff, C. H. W. Hirs & S. N. Timasheff, vol. 115, pp. 41-54. Academic Press Inc., Orlando, Florida.

Howard, E., Sanishvili, R., Cachau, R. E., Mitschler, A., Chevrier, B., Barth, P., Lamour, V., Joachimiak, A., Van Zandt, M., Sibley., E., Moras, D. & Podjarny, A. *The crystal structure of human Aldose Reductase at 0.66 Å: Experimentally observed protonation states and atomic interactions have implications for the inhibitor mechanism*. Submitted to Proteins.

Karle, J. & Hauptmann, H. (1956). *A theory of phase determination for the four types of non-centrosymmetric space groups 1P222, 2P22, 3P12, 3P22*. Acta Cryst. **9**, 635-651.

Karle, J. (1980). *Some Developments in Anomalous Dispersion for the Structural Investigation of Macromolecular Systems in Biology*. Int. J. Quant. Chem. **7**, 357-367.

Kleywegt, G. J. & Jones, T. A. (1996). *xdlMAPMAN and xdlDATAMAN - programs for reformatting, analysis and manipulation of biomacromolecular electron-density maps and reflection datasets*. Acta Cryst **D52**, 826-828.

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

La Fortelle, E. de & Bricogne, G. (1997). *Maximum-Likelihood Heavy-Atom Parameter Refinement for Multiple Isomorphous Replacement and Multiwavelength Anomalous Diffraction Methods*. In *Methods in Enzymology*, edited by C. Carter & R. Sweet, vol. 276, pp. 472-494. Academic Press Inc., Orlando, Florida.

Lamour, V., Barth, P., Rogniaux, H., Poterszman, A., Howard, E., Mitschler, A., Van Dorsselaer, A., Podjarny, A. & Moras, D. (1999) *Production of crystals of human aldose reductase with very high resolution diffraction*. Acta Cryst. **D55**, 721-723.

Lee, Y. S., Hodoscek, M., Brooks, B. R. & Kador, P. F. (1998). *Catalytic mechanism of aldose reductase studied by the combined potentials of quantum mechanics and molecular mechanics*. Biophys Chem. **70**, 203-216.

Li, J., Derewenda, U., Dauter, Z., Smith, S. & Derewenda, Z. S. (2000). *Crystal Structure of the Escherichia coli thioesterase II, a homolog of the human NEF binding enzyme*. Nat. Stru. Biol. **7**, 555-559.

Lunin, V. Y. (1988). *Use of the information on electron density distribution in macromolecules*. Acta Cryst. **A44**, 144-150.

Lunin, V. Y. & Woolfson, M. M. (1993). *Mean phase error and the map-correlation coefficient*. Acta Cryst. **D49**, 530-533.

McRee, D. E. (1999). *Xtalview/Xfit – A versatile Program for Manipulating Atomic Coordinates and Electron Density*. J. Struct. Biol. **125**, 156–165.

Meritt, E. A. & Bacon, D. J. (1997). *Raster3D: photorealistic molecular graphics*. In *Methods in Enzymology*, edited by C. Carter & R. Sweet, vol. 277, pp. 505-524. Academic Press Inc., Orlando, Florida.

Miller, R., Gallo, S. M., Khalak, H. G. & Weeks, C. M. (1994). *Snb: Crystal structure determination via shake-and-bake*. J. Appl. Cryst. **27**, 613-621.

Morris R. J. & Bricogne, G. (2003). *Sheldrick's 1.2 Å rule and beyond*. Acta Cryst. **D59**, 615-617.

Murshudov, G. N, Vagin A. A. & Dodson, E. J. (1997). *Refinement of Macromolecular Structures by the Maximum-Likelihood Method*. Acta Cryst. **D53**, 240-255

Otwinowski, Z. & Minor, W. (1997). *Processing of X-Ray Diffraction Data Collected in Oscillation Mode*. In *Methods in Enzymology*, edited by C. Carter & R. Sweet, vol. 276, pp. 307-326. Academic Press Inc., Orlando, Florida.

Pannu, N. S. & Read, R. J. (1996). *Improved structure refinement through maximum likelihood*. Acta Cryst. **A52**, 659-668.

Perrakis, A., Morris, R. J. & Lamzin, V. S.(1999). *Automated protein model building combined with iterative structure refinement*. Nat. Stru. Biol. **6**, 458–463.

Petrash, M. J., Harter, T. M., Devine, C. S., Olins, P. O., Bhatnagar, A., Liu, S. & Srivastava, S. K. (1992). *Involvement of Cysteine Residues in Catalysis and Inhibition of Human Aldose Reductase*. J. Biol. Chem. **267**, 24833-24840.

Ramachandran, G. N. & Sassiekharan, V. (1968). *Conformation of polypeptides & proteins*. Adv. Prot. Chem. **28**, 283-437.

Read, R. J. (1986). *Improved Fourier coefficients for maps using phases from partial structures with errors*. Acta Cryst. **A42**, 146-149.

Read, R. J. (1990). *Structure-factor probabilities for related structures*. Acta Cryst. A**46**, 900-12.

Sayle, R. & Milner-White, E. J. (1995). *Rasmol: Molecular graphics for all.* Trends in Biochemical Sciences **20**, 374.

Schmidt, A., Gonzalez, A., Morris, R. J., Costabel, M., Alzari, P. M. & Lamzin, V. (2002). *Advantages of high-resolution phasing: MAD to atomic resolution*. Acta Cryst. **D58**, 1433-1441.

Schneider, T. R. (2002). *A genetic algorithm for the identification of conformationally invariant regions in protein molecules*. Acta Cryst. **D58**, 195-208.

Schneider, T. R. & Sheldrick, G. M. (2002). *Substructure solution with SHELXD*. Acta Cryst. **D58**, 1772-1779.

Selmer, M., Al-Karadaghi, S., Hirokawa, G., Kaji, A. & Liljas, A. (1999). *Crystal Structure of Thermotoga Maritima Ribosome Recycling Factor: A tRNA Mimic*. Science **286**, 2349-52.

Sheldrick, G. M. & Schneider, T. R. (1997). *SHELXL: High-Resolution Refinement*. In *Methods in Enzymology*, edited by C. Carter & R. Sweet, vol. 277, pp. 319-343. Academic Press Inc., Orlando, Florida.

Sheldrick, G. M., Hauptman, H. A., Weeks, C. M., Miller, R. & Usón, I. (2001). *Ab initio phasing.* In *International Tables for Crystallography Vol. F: Crystallography of Biological Macromolecules*, edited by M. G. Rossmann & E. Anold, pp. 333-345. Kluwer Academic Publishers, Dordrecht, 2001.

Sheldrick, G. M. (2002). *Macromolecular phasing with SHELXE*. Z. Krist. **217**, 644–650.

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

Thaimattam R., Sheldrick, G. M. & Jaskolski, M. *Experimentally-phased maps in atomic-resolution protein crystallography. A case study*. To be published.

Tronrud, D. E. (1992). *Conjugate-direction minimization: An improved method for the refinement of macromolecules*. Acta Cryst. **A48**, 912-916.

Urzhumtsev, A., Tête-Favier, F., Mitschler, A., Barbanton, J., Barth, P., Urzhumtseva, L., Biellmann, J.-F., Podjarny, A. D. & Moras, D. (1997). *A 'specificity' pocket inferred from the crystal structures of the complexes of aldose reductase with the pharmaceutically important inhibitors tolrestat and sorbinil*. Structure **5**, 601-612.

Usón, I. & Sheldrick, G. M. (1999). *Advances in Direct Methods for Protein Crystallography*. Curr. Opin. Struct. Biol. **9**, 643-648.

Vellieux, F. M. D. & Read, R. J. (1997). *Noncrystallographic Symmetry Averaging in Phase Refinement and Extension*. In Methods in Enzymology, edited by C. Carter & R. Sweet, vol. 277, pp. 18-52. Academic Press Inc., Orlando, Florida.

Wallace, A. C., Laskowski, R. A. & Thornton, J. M. (1995). *LIGPLOT: A program to generate schematic diagrams of protein-ligand interactions. Prot. Eng*. **8**, 127-134.

Walsh, M. A., Otwinowski, Z., Perrakis, A., Anderson, P. M. & Joachimiak, A. (2000). *Structure of cyanase reveals that a novel dimeric and decameric arrangement of subunits is required for formation of the enzyme active site*. Structure **8**, 505-514.

Wang, B. C. (1985). *Resolution of Phase Ambiguity in Macromolecular Crystallography*. In *Methods in Enzymology*, edited by H. W. Wyckoff, C. H. W. Hirs & S. N. Timasheff, vol. 115, pp. 90-112. Academic Press Inc., Orlando, Florida.

Weeks, C. M. & Miller, R. (1999). *The design and implementation of SnB v2.0*. J. Appl. Cryst. **32**, 120–124.

Wimberly, B. T., Brodersen, D. E., Clemons, W. M., Morgan-Warren, R. J., Carter, A. P., Vonrhein, C., Hartsch, T. & Ramakrishnan V. (2000). *Structure of the 30S ribosomal subunit*. Nature **407**, 327-339.

Yang, C. & Pflugrath, J. W. (2001). *Applications of anomalous scattering from S atoms for improved phasing of protein diffraction data collected at Cu Ka wavelength*. Acta Cryst. **D57**, 1480-1490.

Appendix A –Restraint Violations in the refinement of hAR2

Table 1: FLAT restraint violations sorted in descending order for the sigma units. The planarity target value is 0.000, the deviations from planarity are given without +/- sign. 1 Sigma = 0.1 for phenyl rings (Phe, Tyr, Lig), Trp ring systems and guanide groups (Arg). 1 Sigma = 0.5 for amide groups (two residue numbers given).

Residue(s)	dev. fr. zero	Sigma units	Residue(s)	dev. fr. zero	Sigma units
Tyr209	1.6782	16.7820	Lig320	0.4266	4.2660
Trp79	0.9452	9.4520	297/298	2.0227	4.0454
Tyr39	0.9217	9.2170	Arg40	0.3963	3.9630
Tyr48	0.8409	8.4090	Tyr309	0.3922	3.9220
Tyr107	0.7410	7.4100	181/182	1.9505	3.9010
Arg255	0.6430	6.4300	108/109	1.7838	3.5676
Arg69	0.6365	6.3650	311/312	1.6736	3.3472
Trp111	0.5891	5.8910	257/258	1.6338	3.2676
Trp20	0.5585	5.5850	296/297	1.6160	3.2320
132/133(a)	2.3471	4.6942	309/310	1.5901	3.1802
132/133(b)	2.3194	4.6388	262/263	1.5582	3.1164
Trp141	0.4497	4.4970	Arg163	0.3116	3.1160

Table 2: CHIV restraint violations sorted in descending order. The target volume is 0.000 for proline nitrogen atoms, i.e. the geometry is restrained to be planar. 1 sigma = 0.1 for all CHIV restraints.

Residue (Atom)	Chiral Volume	Target	Deviation	Sigma units
Pro310 (N)	0.7294	0.0000	0.7294	7.294
Pro218 (N)	0.6258	0.0000	0.6258	6.258
Pro188 (N)	0.5333	0.0000	0.5333	5.333
Pro117 (N)	0.4642	0.0000	0.4642	4.642
Pro261 (N)	0.4180	0.0000	0.4180	4.180
Asp102 (Ca)	2.1148	2.5030	0.3882	3.882
Pro112 (N)	0.3797	0.0000	0.3797	3.797
Pro252 (N)	0.3469	0.0000	0.3469	3.469
Pro215 (N)	0.3343	0.0000	0.3343	3.343
Val37 (Ca)	2.2080	2.5160	0.3080	3.080

Table 3: BUMP restraint violations sorted in descending order. 1 sigma = 0.02 for all BUMP restraints.

Atoms	Distance	Target	Deviation	Sigma
C_76 CG_77	3.1608	3.3000	0.1392	6.9600
H20_318 H12B_318	1.9631	2.1000	0.1369	6.8450
CD1_209 C_209	3.1707	3.3000	0.1293	6.4650
HB2_296a HG2_296a	1.9796	2.1000	0.1204	6.0200
CD2_41 C_41	3.1870	3.3000	0.1130	5.6500
CD1_209 C17_318	3.1907	3.3000	0.1093	5.4650
H0A_84b HG3_84b	1.9934	2.1000	0.1066	5.3300
O_83a CG2_135a	2.6952	2.8000	0.1048	5.2400
CG_68b OD2_134b	2.6993	2.8000	0.1007	5.0350
HG2_296a HD2_296a	2.0019	2.1000	0.0981	4.9050
CD_293b O_4026b	2.7045	2.8000	0.0955	4.7750
CG_79 N_80	2.9067	3.0000	0.0933	4.6650
HA_61 HD1_61	2.0086	2.1000	0.0914	4.5700
O_213 C_214	2.7130	2.8000	0.0870	4.3500
CD_262 C_262	3.2156	3.3000	0.0844	4.2200
NE2_59b O_4046b	2.4173	2.5000	0.0827	4.1350
C_152 CG1_153	3.2192	3.3000	0.0808	4.0400
H0_135a HG2B_135a	2.0196	2.1000	0.0804	4.0200
CD_146b NH1_293b	2.9265	3.0000	0.0735	3.6750
O_216 C_217	2.7272	2.8000	0.0728	3.6400
HD1_69 HH1B_69	2.0281	2.1000	0.0719	3.5950
O_84b C_85	2.7283	2.8000	0.0717	3.5850
HA_138 HD2A_138	2.0290	2.1000	0.0710	3.5500
CD1_152b O_4042b	2.7323	2.8000	0.0677	3.3850
CB_251 CD_252	3.2347	3.3000	0.0653	3.2650
HE1_293b HH1C_293b	2.0359	2.1000	0.0641	3.2050
CE2_209 C21_318	3.2362	3.3000	0.0638	3.1900
HG1_183 HD2_209	2.0399	2.1000	0.0601	3.0050

Table 4: SIMU restraint violations sorted in descending order for the sigma units. 1 sigma = 0.2 for the first line, otherwise 0.01. The target value is always zero.

residue(s)	ADP / atoms	deviation from zero	sigma units
Lys194	U33 CE NZ (conf. A)	0.7611	3.8055
Asp277	U33 CB CG (conf. A)	0.3460	3.4600
Gln59	U33 CG CD (conf. A)	0.3302	3.3020
Met168	U11 CG Se (conf. B)	0.3125	3.1250
Wat3017 / Wat4034	U22 O O (conf. A)	0.3041	3.0410

Table 5: DELU restraint violations sorted in descending order for the sigma units. 1 sigma = 0.01. The target value is always zero.

residue	atoms	deviation from zero	sigma units
Cit321	C3-O5 (conf. B)	0.0436	4.3600
Lys100	CD-CE	0.0388	3.8800
Lys100	CG-CE	0.0379	3.7900
271	CA-CG (conf. A)	0.0352	3.5200
Cit321	C6-O5 (conf. B)	0.0350	3.5000
26	CG-OE1 (conf. A)	0.0308	3.0800

Table 6: ISOR restraint violations by water oxygen atoms sorted in descending order for the sigma units. To shorten the list only values greater than 4 sigma units are listed. There are altogether 60 violations over 3 sigma. 1 sigma = 0.1. The target value is always zero.

residue	ADP	deviation from zero	sigma units
Wat2617	U12	0.4534	4.5340
Wat2450	U12	0.4442	4.4420
Wat2441	U22	0.4401	4.4010
Wat2611	U23	0.4395	4.3950
Wat2379	U12	0.4326	4.3260
Wat2330	U23	0.4154	4.1540
Wat2223	U23	0.4150	4.1500
Wat2397	U11	0.4098	4.0980
Wat2397	U12	0.4075	4.0750
Wat2490	U12	0.4042	4.0420
Wat2330	U11	0.4029	4.0290

Table 7: DFIX restraint violations sorted in descending order for sigma units. Deviations are given as absolute values. One Sigma is 0.02

Residue(s)	Atoms	Distance	Target	Deviation	Sigma units
Arg296	Cb – Cg (conf. A)	1.3871	1.5200	0.1329	6.6450
Arg296	Cb – Cg (conf. B)	1.4123	1.5200	0.1077	5.3850
Lys100	CB – CG	1.4242	1.5200	0.0958	4.7900
134 / 135	C – N (conf. A)	1.2353	1.3290	0.0937	4.6850
Pro231	CB – CG	1.4056	1.4920	0.0864	4.3200
83 / 84	C – N (conf. B)	1.2507	1.3290	0.0783	3.9150
Met168	Se – Ce (conf. B)	1.8532	1.9300	0.0768	3.8400
Met168	Se – Ce (conf. A)	2.0051	1.9300	0.0751	3.7550
83 / 84	C – N (conf. A)	1.2586	1.3290	0.0704	3.5200
Arg3	CD – NE	1.3946	1.4600	0.0654	3.2700
Glu146	CG – Cd (conf. B)	1.4531	1.5160	0.0629	3.1450

Table 8: DANG restraint violations sorted in descending order for sigma units. Deviations are given as absolute values. One sigma is 0.04.

Residue(s)	Atoms	1,3-distance	target	deviation	sigma units
Thr135	Cg2 – Og1 (conf. a)	2.6612	2.3730	0.2882	7.2050
Cys298	CB – N (conf. a)	2.1913	2.4550	0.2637	6.5925
Cys298	CB – N (conf. b)	2.7093	2.4550	0.2543	6.3575
Gln59	OE1 – NE2 (conf. b)	2.0158	2.2450	0.2292	5.7300
Arg293	CA – Cg (conf. A)	2.3462	2.5590	0.2128	5.3200
His84	N – C (conf. a)	2.2725	2.4620	0.1895	4.7375
Glu267	CB – Cd (conf. A)	2.7137	2.5260	0.1877	4.6925
Arg296	C – Cb (conf. A)	2.3290	2.5040	0.1750	4.3750
Ile169	CA – CG1 (conf. B)	2.4014	2.5760	0.1746	4.3650
134 / 135	O – N (conf. b)	2.0779	2.2500	0.1721	4.3025
Glu267	CG – OE1 (conf. A)	2.5440	2.3790	0.1650	4.1250
Val297	CA – N (conf. b)	2.5850	2.4250	0.1600	4.0000
Gln59	CG – OE1 (conf. b)	2.5521	2.3930	0.1591	3.9775
Glu279	CB – CD (conf. a)	2.6848	2.5260	0.1588	3.9700
Lys307	CD – NZ	2.6440	2.4930	0.1510	3.7750
Cys298	CA – SG (conf. a)	2.6591	2.8100	0.1509	3.7725
Glu126	CB – CD (conf. A)	2.6765	2.5260	0.1505	3.7625

Gln197	CA – CG (conf. B)	2.4125	2.5590	0.1465	3.6625
196 / 197	C – Ca (conf. b)	2.2920	2.4350	0.1430	3.5750
Lys119	CD – NZ (conf. A)	2.6352	2.4930	0.1422	3.5550
Glu271	CA – CG (conf. A)	2.7005	2.5590	0.1415	3.5375
Cys298	C – CB (conf. a)	2.6439	2.5040	0.1399	3.4975
Glu271	CG – OE1 (conf. a)	2.2404	2.3790	0.1386	3.4650
His83	O – N (conf. b)	2.1127	2.2500	0.1373	3.4325
Lys100	CA – CG	2.6959	2.5590	0.1369	3.4225
83 / 84	O – N (conf. a)	2.1184	2.2500	0.1316	3.2900
Lig318	C29 – C30	2.2179	2.3490	0.1311	3.2775
Arg296	CB – CD (conf. B)	2.3789	2.5100	0.1311	3.2775
Lys242	CG – CE (conf. A)	2.6373	2.5100	0.1273	3.1825
Met253	CG – CE (conf. A)	3.0053	2.8810	0.1243	3.1075
Val297	CA – N (conf. a)	2.3008	2.4250	0.1242	3.1050
Arg293	CA – CG (conf. B)	2.6830	2.5590	0.1240	3.1000
Glu271	CB – N (conf. A)	2.5781	2.4550	0.1231	3.0775
Arg293	CB – CD (conf. A)	2.3889	2.5100	0.1211	3.0275
Glu29	CG – OE2 (conf. b)	2.2587	2.3790	0.1203	3.0075

Appendix B – Free Variables in the refinement of hAR2

Table 1: The assignment of free occupancy variables and their refined values. Missing variable numbers are related to variables temporarily used for threefold disorder (SUMP), and were not used later.

variable #	value	used for residues
2	0.5914	71 83 84 133 134 135 136 137 6101 6102 6103 6104 6105 6106
3	0.5365	29 126 127 6301 6302 6303 6304 6305
4	0.5100	68 70 6501 6502 6503
5	0.7127	117 6901
6	0.6970	60 63 64 6951 6952 6953
7	0.7670	59 6911
8	0.5925	67
9	0.5386	93 97 178 6601 6602 6603 6604 6605 6606
10	0.6050	162 163 164 194 321 6201 6202 6203 6204
11	0.8959	12
12	0.9536	144
13	0.8736	253
14	0.9280	285
15	0.6336	102 6991 6992
16	0.6022	120 6981
17	0.5360	146 293 6921 6922 6923
18	0.6738	152
19	0.5795	154
20	0.6575	26 7021
21	0.5839	239 6931
22	0.6241	241
26	0.7384	305 6961 6962 6963
27	0.6483	1001
31	0.7619	129
32	0.7288	119
33	0.6470	169
34	0.5358	193 6001 6002 6003 6004 6005 6006
35	0.6080	267 6941
36	0.6521	271 6801 6802 6803
37	0.5744	279 6701 6702 6703
41	0.5694	225 295 296 297 298 299 300
42	0.6499	116

43	0.6355	311 312 313 7011
44	0.5849	179
45	0.6510	217
46	0.6894	242 277
47	0.5403	319 6401 6402 6403 6404 6405
48	0.7145	229 7001
49	0.5211	234
50	0.7385	256
51	0.5266	321
52	0.6568	197 6971 6972 6973
53	0.7192	40 41
54	0.7399	5001
55	0.5972	5002
56	0.6896	5003
57	0.6966	5004
58	0.6573	5005
59	0.6745	5006
60	0.7424	5007
61	0.6000	85

Danksagung

An erster Stelle gebührt Dank meinem Doktorvater, Prof. George M. Sheldrick Ph. D., für die interessante Aufgabenstellung meiner Promotion und die stete Hilfs- und Diskussionsbereitschaft während meiner gesamten bisherigen Tätigkeit am Lehrstuhl für Strukturchemie, einschließlich meiner Diplomarbeit.

Ganz herzlich danke ich meinem Betreuer Dr. Thomas R. Schneider für die angenehme und erfolgreiche Zusammenarbeit, für seine immer freundliche und aufmunternde Art, und vor allem für die wertvollen Ideen und Ratschläge bezüglich meiner Promotions-Projekte.

Bei der Kommission der Europäischen Union bedanke ich mich für die Finanzierung meiner Promotion im Rahmen des „*Autostruct*"-Projektes, EU-Vertrag *QLRI-CT-2000-00398*.

Herrn Prof. Alberto Podjarny vom IGBMC Strasbourg möchte ich für die erfolgreiche Kollaboration im Aldose-Reduktase-Projekt und die Bereitstellung der hervorragenden kristallographischen Daten der Aldose-Reduktase danken.

Ich danke allen Mitgliedern unserer Abteilung für die angenehme Arbeitsatmosphäre, vor allem aber Frau Dipl.-Chem. Eftichia Alexopoulos und Frau Dipl.-Chem. Ilka Müller für das besonders nette Arbeitsverhältnis und die stete freundschaftliche Unterstützung über die Arbeit hinaus. Für das Korrekturlesen meiner Doktorarbeit bin ich beiden zu besonderem Dank verpflichtet. Auch bei Herrn Dipl.-Chem. Jose Antonio Cuesta-Seijo möchte ich mich für das Korrekturlesen bedanken.

Schließlich danke ich meinen Eltern für die finanzielle Ermöglichung meines Studiums und für alles, was sie mir in meinem Leben an Liebe, Rat und Unterstützung gegeben haben.

Publikationen

Diedrich, F., Klingebiel, U., Dall'Antonia, F., Lehmann, C., Noltemeyer, M. & Schneider, T. R. (2000). *Asymmetric Tris- and Cyclic Silylhydroxylamines from Trimeric and Tetrameric Lithium-N,N-Bis(silyl)hydroxylamides*, Organometallics **19**, 5376-5383.

Gellermann, E., Klingebiel, U., Pape, T., Dall'Antonia, F., Schneider, T. R. & Schmatz, S. (2001). *Silylhydrazine und dimere N,N'-Dilithium-N,N'-bis(silyl)hydrazide - Synthesen, Reaktionen, Isomerisierungen*. Z. anorg. allg. Chem. **627**, 2581-2588.

Bertasso, M., Holzenkämpfer, M., Zeeck, A., Dall’Antonia, F. & Fiedler, H.-P. (2001). *Bagremycin A and B, Novel Antibiotics from Streptomyces sp. Tü 4128*. J. Antibiot. **54**, 730-736.

Most, K., Köpke, S., Dall'Antonia, F. & Mösch-Zanetti, N. C. (2002). *The first molybdenum dioxo compounds with* η^2*-pyrazolate ligands: crystal structure and oxo transfer properties*. Chem. Commun., 1676-1677.

Dall'Antonia, F., Baker, P. J. & Schneider, T. R. (2003). *Optimization of Selenium substructures as obtained from SHELXD*. Acta Cryst. **D59**, 1987-1994.

Dall’Antonia, F., Sheldrick, G. M., Howard, E., Hazemann, I., Petrova, T., Mitschler, A., Sanishvili, R., Joachimiak, A., Moras, D., Podjarny, A. & Schneider, T. R. *Validation of Multiple Conformations in Experimental Electron Density Maps of human Aldose Reductase*, in preparation.

Lebenslauf

Name:	Fabio Dall’Antonia
Geburtsdatum, -ort:	9. Oktober 1974 in Göttingen
Eltern:	Romeo Dall’Antonia und Astrid Dall’Antonia, geb. Herrmann
Staatsangehörigkeit:	deutsch und italienisch
Familienstand:	ledig

1981 – 1982	Leineberg-Grundschule in Göttingen
1982 – 1985	Hagenberg-Grundschule in Göttingen
1985 – 1987	Orientierungsstufe Bert-Brecht-Schule in Göttingen
1987 – 1994	Felix-Klein-Gymnasium in Göttingen
Juni 1994	Abitur
Okt. 94 – Sep. 95	Zivildienst an der Hainberg-Klinik in Göttingen
Okt. 95 – Okt. 97	Grundstudium der Chemie
Oktober 1997	Vordiplom in Chemie
Okt. 97 – Jan. 99	Hauptstudium der Chemie
Feb. 99 – Mai 2000	Diplomarbeit zum Thema „Röntgenstrukturuntersuchungen an siliziumorganischen Verbindungen und Chinolon-Derivaten und Strukturverfeinerung des α-Amylase-Inhibitors Tendamistat“ am Lehrstuhl für Strukturchemie bei Prof. Sheldrick Ph. D.
Mai 2000	Diplom in Chemie
Jul. 2000 – Sep. 03	Dissertation zum Thema „Studies on the Crystallographic Phasing of Proteins: Substructure Validation and MAD-phased Electron Density Maps at Atomic Resolution“ am Lehrstuhl für Strukturchemie bei Prof. Sheldrick Ph. D.