Practical approaches to macromolecular X-ray structure determination

D i s s e r t a t i o n
zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
"Doctor rerum naturalium"
der Georg-August-Universität Göttingen

vorgelegt von
Andrea Regina Shirin Thorn

Göttingen, 2011

Referent: Prof. George M. Sheldrick FRS
Koreferent: Dr. Birger Dittrich

Tag der mündlichen Prüfung: 23. Juni 2011

Für meine Familie,
im Gedenken an
Dr. Marianne Thorn (1921 –2011)

Contents

Somerville College May 17th, 1931

My dearest Mummy and Daddy,
(...)
A few days ago Dr. Joseph wrote to me to say that he had asked Professor Lowry about the possibility of my doing X-ray work on crystals – and whether it was a good thing. (...) And all that sounded very nice – really excellent just then – since the X-ray work would be useful in absolutely anything I decided to do ever afterwards and yet if I did not do it now – I probably should not have the chance again. But at the moment I'm feeling quite appalled at the prospect. There will be such a fearful lot of work – and mathematics – involved. And I was just beginning to rejoice so much in the idea of a nice quiet organic research that would involve no brain whatsoever. As it is, it will be pure brain work – I'm just shivering in my shoes – terribly afraid I really am trying to force too much on one poor little brain that is almost non-existent already.
(...)
Of course, if I can really do it it will be rather priceless...

– Dorothy Crowfoot Hodgkin (1910 – 1994)

Introduction

By the time Dorothy Hodgkin endeavoured to do her PhD about X-ray work on crystals, not much was known about the role of macromolecules in life. Proteins were believed to be globules or micelles with unknown structure. Five years earlier, it had been shown by Sumner (1926) that the urease – as an enzyme – was a protein. And only in 1937, Astbury found X-ray diffraction patterns which proved the repetitive structure of DNA (Astbury, 1947). The X-ray structure determination of biological macromolecules in the following decades therefore was a revolution, giving way to the field of molecular biology and to biochemistry as known today. Still, macromolecular X-ray structures are unsurpassed in precision and detail.

But these structures cannot directly be derived from X-ray diffraction data (Rodríguez *et al.*, 2009). Phases and a molecular model are needed to interprete the measured reflections, and hence, understanding of the underlying principles is crucial. This understanding is the driving force behind this work on macromolecular X-ray structure determination.

New practical approaches for phasing are given alongside with two protein structures obtained by this means. ANODE, a tool for the evaluation of and validation by experimental X-ray data is presented and a new atomic displacement restraint for the refinement of macromolecular models is given.

1 Background

1.1 Phasing methods

Each reflection hkl is related to a structure factor F_{hkl}, which can be written as a complex number. It is composed of scattering factors f from every atom in the unit cell. If we know the amplitude and the phase of each structure factor exactly, we can calculate a perfect distribution of electron density in the asymmetric unit (ASU) using the Fourier transform. The amplitude $|F_{hkl}|$ is related to the reflection intensity by $|F_{hkl}|^2 \propto I_{hkl}$, but the phase ϕ_{hkl} cannot be measured directly. This is called the crystallographic phase problem. Several methods exist to solve it. For small molecule structures, Patterson and direct methods are the most common; for macromolecules, only few structures can be solved by direct methods. They are commonly solved by molecular replacement or by experimental phasing methods. Density modification can be applied to yield additional phase information (Rupp, 2009).

1.1.1 Molecular replacement

Molecular replacement (MR) phasing is often employed for similar structures like mutants, co-crystallization experiments and in polymorphism. It is also applied where experimental phasing is not feasible. For MR, a search model similar to the target structure is needed.

The number of potential search models increases with the size of the protein data bank (Berman *et al.*, 2002), and new applications even generate homology models specifically designed for MR (for example, Claude *et al.* 2004). Nonetheless, finding a good model can be challenging. Chothia & Lesk (1986) found that $\sigma_r \approx 0.4 \cdot exp\left[1.87 \cdot (1 - s)\right]$ where σ_r is the r.m.s. coordinate deviation and s the sequence identity. A good search model should have no greater r.m.s. coordinate deviation than 1.5 Å, and consequently, the sequence identity should be more than 30%. Trimming to the most conserved and rigid parts of the structure (for example the main chain) can lower the r.m.s. deviation and hence improve the chance of successful phasing.

The search probe is located and oriented in the unit cell at the same position as the measured structure to achieve starting phases ϕ_{hkl}. In most MR programs, the positioning problem is broken down into two steps: A three-dimensional orientation search and a three-dimensional translation search.

Rotation solutions are scored using *Patterson maps*. These maps represent the *Patterson function*, which is the Fourier transform of $|F_{hkl}|$ on a plane or in space. A cross rotation search (Rupp, 2009) based on the overlapping of *Patterson maps* is conducted, matching interatomic vectors.

As small errors in the rotation search can prevent finding a suitable translation solution, the angular increments used for the cross rotation search have to be reasonable. To find the orientation faster, the *Patterson function* can be replaced with its Fourier transform. This is called

"fast rotation function". Non-crystallographic symmetry (NCS) is found by *Patterson self-rotation search*; known NCS can be used to limit the rotational search space ("locked rotation search").

The translation search also makes use of *Patterson maps* by locating the position of the model in the Cheshire cell, which is the space between potential unit cell origins at a given state of the search. As for rotation searches, "fast translation functions" can be used. Steric overlap penalty functions further improve the solution search (Harada *et al.*, 1981). As the best scoring rotation solution isn't always correct, it is better to score rotation solutions against translation functions.

A general drawback in molecular replacement is model bias. Other than in experimental phasing, the phase information is biased by the model. This is especially true at low resolution, where the data-to-parameter ratio is low (Rupp, 2009). The map will reflect model features and bias the final structure. These problem can be overcome by MR-SAD, as discussed below.

1.1.2 Experimental phasing

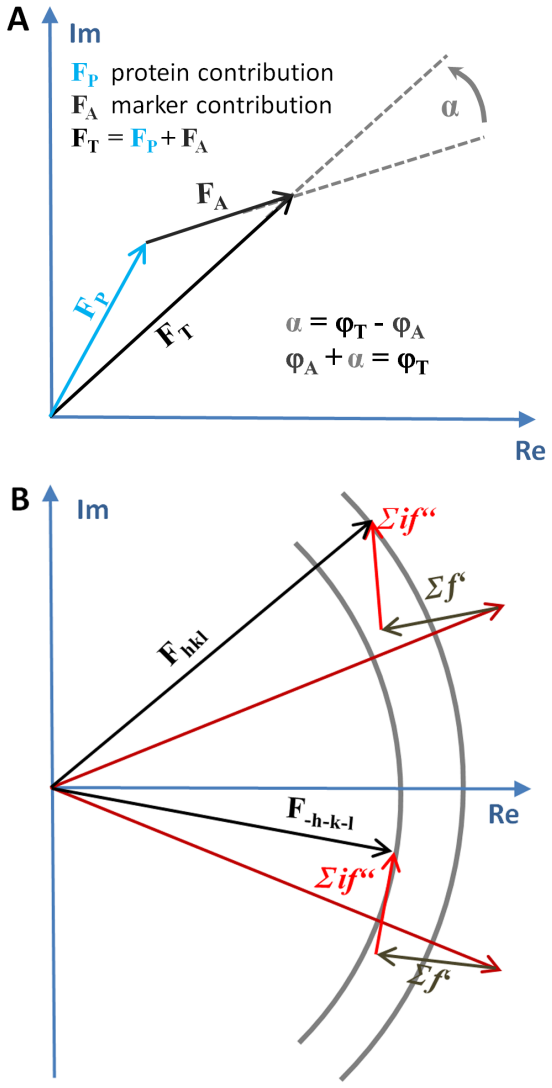

Figure 1.1: **A.** Definition of α. **B.** The contribution if'' breaks Friedel's law.

Experimental phasing methods are – as opposed to MR – independent of a search model and are based on the electronic differences of heavy atoms or anomalously scattering atoms (called anomalous scatterers). These marker atoms form a substructure, which is represented by differences between data sets, or, in case of anomalous scattering, within the same data set (Rupp, 2009). These differences are used to find the marker atom positions, from which starting phases for the macromolecule can be gained by solving the phase equations.

The nature of the marker atom substructure is dependent on the experimental phasing method used. The most common experimental phasing methods are: Isomorphous replacement with heavy atoms, SIR or MIR; radiation-induced phasing (RIP) which uses radiation damage to the substructure but is otherwise similar to SIR; anomalous diffraction methods, SAD and MAD, which use anomalous scatterers and finally SIRAS, which is isomorphous replacement with anomalous scattering (Rupp, 2009).

Native sulfur-based SAD (S-SAD) is a special case of SAD, where the protein's own sulfur is used for phasing. The data measurement has to be precise and a high multiplicity is needed.

Multiple wavelength anomalous diffraction (MAD) can theoretically give a perfect solu-

tion, as it provides orthogonal dispersive and anomalous differences from the same crystal and the two-fold phase ambiguity can be resolved directly from the phasing equations given below.

All of these methods employ the phasing equations, based on the angle α. Each structure factor F_{hkl} (or F_T) in an experimental phasing data set is composed of a protein contribution F_P and a marker atom contribution F_A. The difference between the phases of F_A and F_T is α (see Fig. 1.1).

$$F_T = F_P + F_A$$

$$\alpha = \phi_T - \phi_A$$

If the marker atom positions are known, their contribution F_A can be calculated, including their phase ϕ_A. If α is also known, ϕ_T can be calculated and the phase problem is solved.

Near the absorption edge of an element contained in the measured crystal, significant deviations from Friedel's law ($|F_{hkl}| = |F_{-h-k-l}|$; $\phi_{hkl} = -\phi_{-h-k-l}$) can be observed. These result from resonance with electronic transitions in the atom. The atomic scattering factor f is composed of f_0 which solely depends on the scattering angle θ, the real component $f'$and the imaginary component if'' which are dependent on the X-ray wavelength λ. The contribution if'' breaks Friedel's law, as shown in Fig. 1.1. $|F_{hkl}| - |F_{-h-k-l}|$ is called the Bijvoet difference. The phasing equations (Karle, 1980; Hendrickson *et al.*, 1985) link the amplitudes of reflection hkl and -h-k-l with this effect:

$$|F_{hkl}| = |F_T|^2 + a \cdot |F_A|^2 + b \cdot |F_T| \cdot |F_A| \cdot cos\alpha + c \cdot |F_T| \cdot |F_A| \cdot sin\alpha$$

$$|F_{-h-k-l}| = |F_T|^2 + a \cdot |F_A|^2 + b \cdot |F_T| \cdot |F_A| \cdot cos\alpha - c \cdot |F_T| \cdot |F_A| \cdot sin\alpha$$

$$a = \frac{f''^2 + f'^2}{f_0^2} \quad b = \frac{2f'}{f_0^2} \quad c = \frac{f''^2}{f_0^2} \quad \alpha = \phi_T - \phi_A$$

For each wavelength at which a data set was measured, we have different a, b, c values and two observations (I_{hkl} and I_{-h-k-l}). $|F_A|$, $|F_T|$ and α are unknown. In MAD, data sets from at least two wavelengths can be used to calculate values for the α angle. In SAD, however, only one data set gives us only two observables. We have to make the approximation $|F_T| = 0.5 \cdot (|F_{hkl}| + |F_{-h-k-l}|)$ and get $|F_{hkl}| - |F_{-h-k-l}| = c \cdot |F_A| \cdot sin\alpha$. By using normalized structure factor amplitudes (see page 9), c becomes obsolete. The angle α can be estimated as shown in Fig. 1.2. Hence, we can estimate $|F_A|$ and solve the phasing equations. An inherent two-fold phase ambiguity remains from the α angle estimation, and it can not be distinguished which enantiomorph of the marker atom substructure is correct. Density modification based on disordered solvent regions in the crystal resolve the two-fold phase ambiguity.

1.1.3 MR-SAD

Molecular replacement can be combined in various ways with SAD to amplify weak signals which, taken separately, would not be sufficient for structure solution (Roversi *et al.*, 2010; Roeser

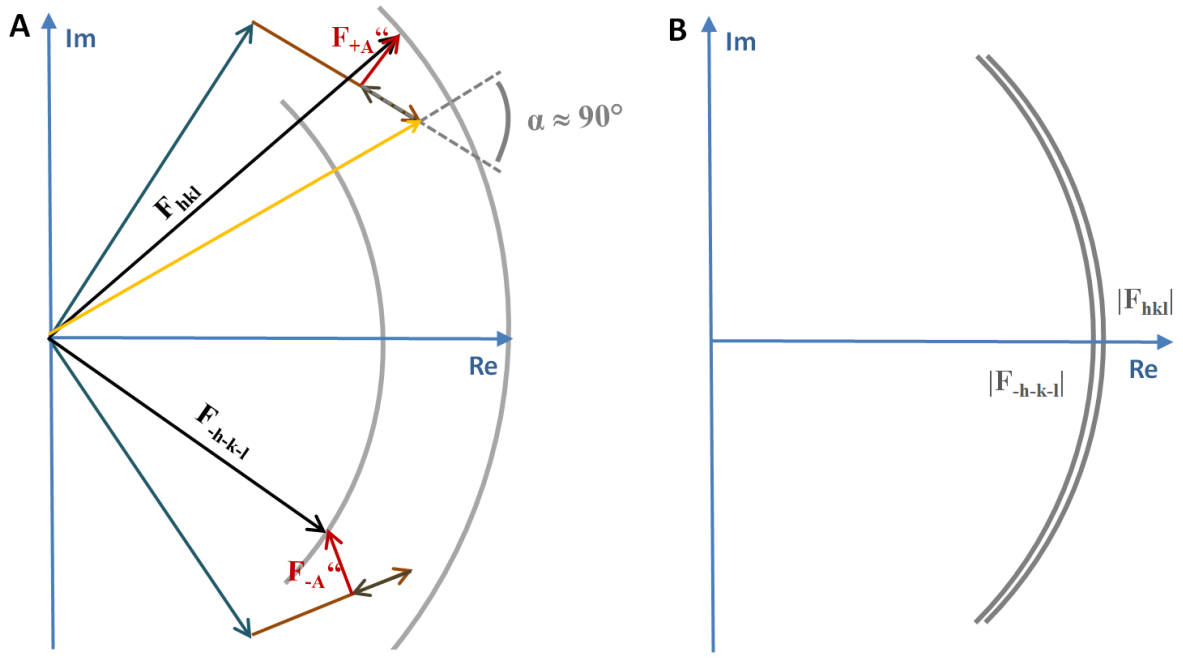

Figure 1.2: Estimation of the angle α for SAD phasing. **A.** $|F_{-h-k-l}|$ is much bigger than $|F_{hkl}|$, marked by grey circles. F_{+A}" has to point in the same direction as F_{hkl}; F_{-A}" in the opposite direction of F_{-h-k-l}. Consequently, α must be close to 90°. If $|F_{-h-k-l}|$ is much smaller than $|F_{hkl}|$, α must be close to 270°. **B.** If $|F_{-h-k-l}|$ is approximately $|F_{hkl}|$, F_{+A}" and F_{-A}" must have small amplitudes or be almost perpendicular to F_{hkl} or F_{-h-k-l}, respectively. The angle α must be close to 0° or 180°. If 0° or 180° apply, it cannot be determined at this stage - a two-fold phase ambiguity results, which can later be solved by density modification.

et al., 2005). This phasing method has been named MR-SAD (Schuermann & Tanner, 2003), and is employed as follows:

A partial or potential MR solution serves as starting point. An anomalous electron density map, eventually with solvent flattening applied, is produced from the placed search model and the anomalous signal in the data. The peaks in this map then hint at the anomalous scatterer positions. Alternatively, the anomalous scatterer positions for native sulfur can be derived directly from the MR solution's cysteine or methionine positions.

In both cases, if the substructure search is successfully "bootstrapped" with these positions, the MR solution has been validated, and phases can be gained as they would be in normal SAD phasing. The resulting electron density is unbiased by the initial MR model.

Even in cases where the anomalous signal is too noisy or weak for conventional SAD, the data might be useful for MR-SAD phasing. The problem of enantiomorph ambiguity is skipped as well, as the MR solution already gives only one possible hand (Panjikar *et al.*, 2009).

1.2 Phasing programs

1.2.1 PHASER

The program PHASER (McCoy *et al.*, 2007) for macromolecular replacement uses *maximum likelihood* target functions to better distinguish between noise and good solutions. Rotation and translation search are separated, and if several models should be placed in the ASU, one is positioned after the other. As input, a **PDB** model or an ensemble of alternative models can be used and a great number of options allow fine-tuning the search for a phase solution.

The program gives the coordinates of positioned models, an MTZ file with phases (and data) and several quality indicators. Among them are the log likelihood gain (LLG), the rotation function Z-score (RFZ) and the translation function Z-score (TFZ).

The LLG in PHASER is defined as (Bunkoczi, personal communication):

$$LLG = \sum_{hkl} ln\left[p(|F_{obs}|; model)\right] - \sum_{hkl} ln(p_{wilson}(|F_{obs}|))$$

The term $p(|F_{obs}|; model)$ is the probability of the observed data given the model orientation and location in the ASU; $p_{wilson}(|F_{obs}|)$ is the likelihood score for a random-atom Wilson distribution. The LLG should increase between different stages of the molecular replacement and it should be positive in any case. It can be used to assess the significance of a solution, but as the LLG is dependent on model and data, it cannot be compared between different structures. This is why Z-scores are used.

A Z-score is computed as the difference of the LLG of a particular rotation or translation solution (in standard uncertainties σ) and the mean LLG of a random sample of orientations, divided by the r.m.s.d. of a random sample of LLG from the mean (Einspahr & Weiss, 2011; Collaborative Computational Project, 2011). A Z-score is therefore not a likelihood score, only a measure of how significant a peak is (Bunkoczi, personal communication).

The RFZ values may not give a clear indication for a solution, especially at high resolution or with NCS present. But a high TFZ (> 7–8) usually gives a good indication of a successful phasing. If the Z-score is below, the solution still could be correct, but there is no way to tell it or pick the correct one from a list of possibilities based on the Z-score alone (Bunkoczi, personal communication). A TFZ less than 5 might indicate a false solution. In monoclinic space groups, the translation search for the first search fragment is carried only out over a plane, because these groups are polar. Because of this, the TFZ can be too low (Read *et al.*, 2006).

1.2.2 SHELXC and SHELXD

The programs SHELXC/D/E (Sheldrick, 2008, 2010) are intended for experimental phasing. To eliminate effects which result from measurement at different scattering angles θ, SHELXC/D/E use E-values (normalised structure factors) which correspond to point atoms instead of atoms with an electron distribution (and atomic displacement).

The normalised structure factor factor amplitudes E are defined as:

$$|E_{hkl}|^2 = \frac{|F_{hkl}|^2/\epsilon}{<|F_{hkl}|^2/\epsilon>}$$

The scale factor ϵ is needed for proper treatment of special position reflections (Massa, 2007). $< |F_{hkl}|^2/\epsilon >$ is the mean calculated per resolution shell. In the case of SAD, the approximation $|F_{hkl}| - |F_{-h-k-l}| = c \cdot |F_A| \cdot sin\alpha$ becomes $|E_{hkl}| - |E_{-h-k-l}| = |F_A| \cdot sin\alpha$; c is dropped.

SHELXC prepares the files for SHELXD and SHELXE. As explained in detail on page on page 8, it estimates α from data or, in case of MAD, RIP etc., calculates them. XPREP (Sheldrick, 2011) has similar functionalities.

SHELXD (Usón & Sheldrick, 1999) locates the marker atom substructure, which it assumes only to consist of one element, so that f' and f'' do not need to be specified. If there are several marker atom types present, this is compensated by occupancies. The program was originally intended for the solution of large small molecule structures by direct methods. Sheldrick's rule (Sheldrick, 1990) states that for direct methods phasing, at least half the number of possible reflections between 1.1 and 1.2 Å resolution have to be well measured, so that atomic electron distributions are separated clearly from each other. Most macromolecular X-ray data extend not to such a high resolution.

SHELXD can be applied, since the substructure data only refers to the marker atoms, which are typically far enough apart from each other to resolve them at medium resolution. Disulphide bridges can be treated as so-called super-sulfurs, and a special option to find these elongated electron density maxima is available (Debreczeni *et al.*, 2003a,b).

The program starts with random marker atom positions or ones which are consistent with the sharpened *Patterson function* ("*Patterson seeding*"). Then a a *dual space* algorithm is employed (Schneider & Sheldrick, 2002), which iterates between two steps:

1. Picking the most promising positions in real space. SHELXD optionally omits 30% of the highest peak positions for the calculation of phases and keeps a given number of positions, which should be as near as possible to the real number of marker atoms in the ASU.

2. Refining the phases in reciprocal space.

After this, marker atom occupancies are refined by two cycles of conjugate gradient least squares (Schneider & Sheldrick, 2002).

1.2.3 SHELXE

SHELXE (Sheldrick, 2008, 2002) calculates ϕ_T, taking the α values into account and generates via Fourier synthesis an initial electron density map. At this stage, the SAD phase angles still have poor quality and are hampered by the two-fold phase ambiguity inherent to SAD and SIR phasing. Density modification improves the phases and resolves the ambiguity.

Density modification in SHELXE is based on regions filled by disordered solvent which has less features than ordered regions of the crystal. Therefore, a high solvent content often gives better results.

While most programs mask solvent regions and then use *solvent flattening*, SHELXE uses the *sphere of influence* algorithm, which exploits that 1,3-distances in macromolecules are often close to 2.42 Å. In this algorithm, the electron density variance on a spherical surface (r = 2.42 Å) around a map voxel is calculated. If the variance is high, this hints to the centre being an actual atomic position. The density of the centre voxel is flipped if negative and optionally sharpened (Sheldrick, 2008). For voxels with low variance, the density is inverted, which after a

few cycles results in flattening. NCS averaging can be used to further improve map quality. The variance of this variance is called contrast in the program's output and is higher for the correct enantiomorph.

The recent SHELXE beta test version (Sheldrick, 2010), which has been extensively tested in this work, also has an option for auto tracing. The iteration between density modification and auto tracing was initially implemented to get a structure solution from a noisy map from poor phases. Later it was used as a general step in experimental phasing. In this work, we expand its use to molecular replacement and MR-SAD. For the auto tracing, potential α helices (if applicable) and tripeptides are searched in the map and extended on their termini. To give a unique trace, they are spliced with regard to the symmetry of the crystal. SHELXE also makes use of a "no-go" mask, which gives areas where existing atoms or symmetry prohibit tracing. The trace is validated by a positive density 2.9 Å from N in the N-H direction (a hydrogen bond donor), a good fit of the trace's atoms in the density, chain length, a relatively good Ramachandran fit (Ramachandran & Sasisekharan, 1968) and a well defined secondary structure – a low variance of ϕ and ψ angles between neighbouring residues.

The command line input to SHELXE usually takes the form of:

`SHELXE XX YY [options]`

`XX` is the file name of the native data (**`HKL`** format). Also, an **`XX.ins`** file is read in for initial phases.

Start phases can also be derived from a file in

`PDB` format: **`SHELXE XX.pda* [options]`**

`PHS` format: **`SHELXE XX.phi* [options]`**

`FCF` format: **`SHELXE XX.fcf [options]`**

`HLC` format: **`SHELXE XX.hlc [options]`**

* **`PDA`** and **`PHI`** file extensions are used so that SHELXE doesn't overwrite the files as it writes out **`PHS`** and **`PDB`**. **`FCF`** and **`PHS`** formats contain data and phases and the electron density can be displayed with them. HLC format files contain Hendrickson-Lattman coefficients.

`YY_fa.hkl` prepared by SHELXC or XPREP and contains $|F_A|$, its uncertainty $\sigma(|F_A|)$ as well as as the phase shift α. Optionally, a **`YY.res`** file with the marker atom positions is read in. If **`YY`** is specified, an **`XX.hat file`** – with the revised marker atom positions – is written out. This can be renamed and reused again, as for example in section 3.4.2 on page 39 of this work.

The following options available for SHELXE beta are used in this work:

syntax	function
-h(N)	the first N marker atoms should be considered
-d(resolution)	resolution cut-off for input data
-e(resolution)	*free lunch* extension (only in the last iteration, if combined with **-a**)
-m(N)	cycles of density modification
-s(fractional)	solvent content
-a(N)	auto tracing iterations
-t(N)	helix and tripeptide search time factor
-q(N)	helix search (in N first iterations)
-n(N)	application of N-fold NCS in auto tracing
-l(N)	space for N $\cdot 10^6$ reflections
-y(resolution)	starting phases from model resolution cut-off
-i	structure inversion (for resolving the two-phase ambiguity)

SHELXE gives different output dependent on options used: If a marker atom list is written out, it will be sorted by the absolute of the anomalous density calculated directly at the peak position. The atoms which are given as "revised" are the ones before the first negative anomalous density value. If poly-Ala tracing was used a **PDB** file with the main chain trace is written.

Finally, SHELXE allows a *free lunch*. In this algorithm, the data are expanded beyond the resolution limit with rough guesses for $|F_{hkl}|$. Missing reflections are completed as well. These additional amplitudes are gained from Fourier transform of the modified electron density map and are normalized to fit an extrapolated Wilson plot. A *free lunch* is typically chosen from 2.0 Å downward. Fourier truncation errors are corrected, as 0 might be a poor estimation for intensities not measured. This only works because the influence of the phases on the electron density map is higher than that of the amplitudes. The map gained by *free lunch* is optimal for initial model building. Such a map is shown in Fig. 1.3 for Hellethionin D.

As quality indicators for a successful tracing, the author gave the criteria of an average poly-Ala chain length of 10 or more as well as a correlation coefficient against native data of 20% or higher (Sheldrick, personal communication). These criteria are not generally applicable, as will be shown in this work.

1.2.4 ARCIMBOLDO

The ARCIMBOLDO *ab-initio* phasing method (Rodríguez *et al.*, 2009) uses α-helical fragments (of 10-14 residues length) as search fragments for the MR program PHASER (McCoy *et al.*, 2007) instead of a particular X-ray structure. As many different small helix position combinations might be correct and PHASER frequently generates several good rotation-translation results, a great many of potential solutions are generated. As the positioned search fragments represent only a fraction of the total structure, it can be difficult to distinguish a correct solution by PHASER quality indicators alone. Therefore, all potential solutions are read into SHELXE, where density modification plus auto tracing are applied to distinguish good solutions and to further improve their phases. ARCIMBOLDO is originally run on a CONDOR computer grid and highly parallelized. To work properly, data extending to at least 2.0 Å is needed – but neither experimental phase information nor a model of the protein are required.

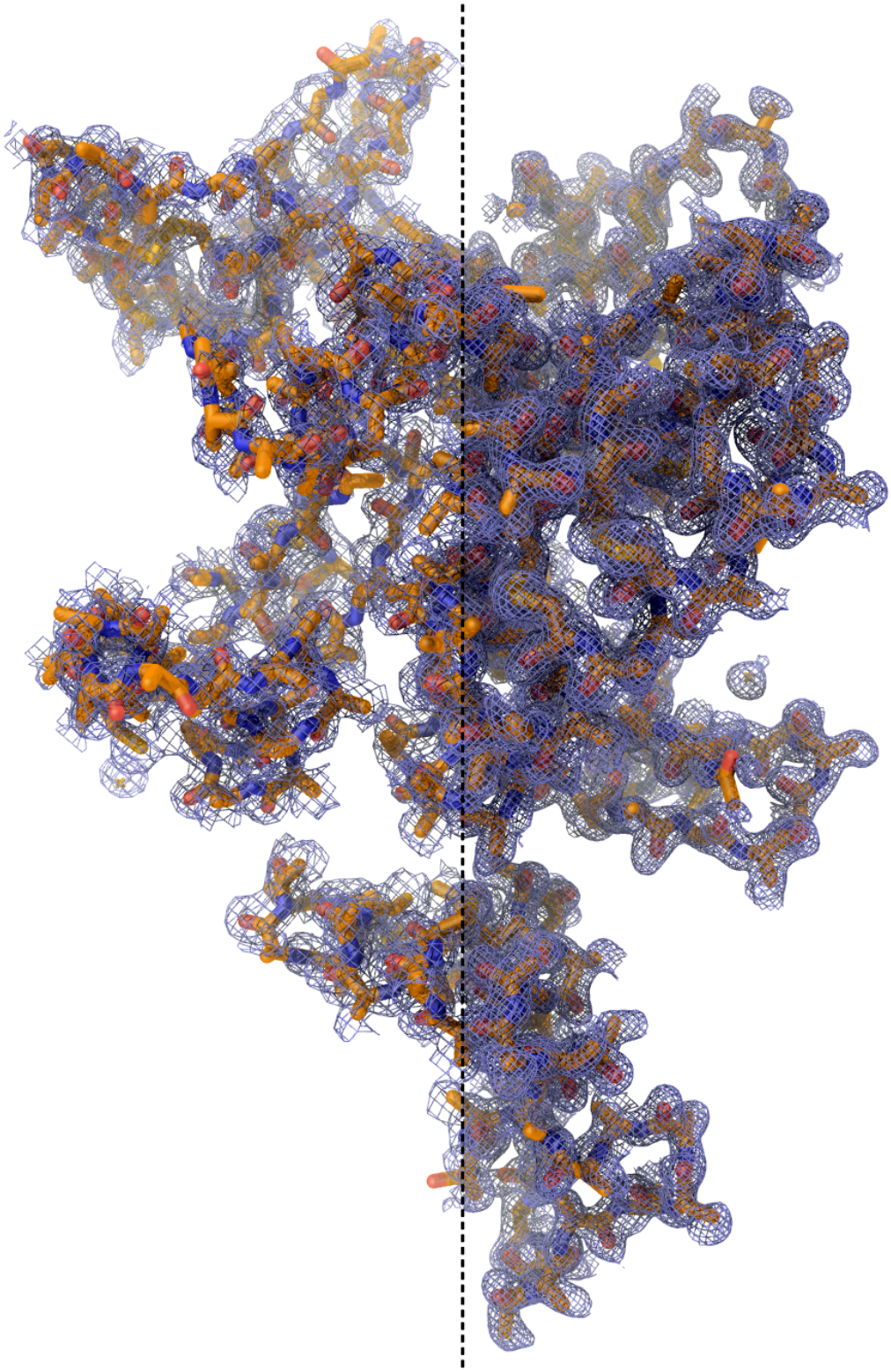

Figure 1.3: *Free lunch* density for Hellethionin D. The data had a resolution of 1.95 Å (top half) and was expanded to 1.0 Å (bottom half).

1.2.5 A note on anisotropic scaling

Crystal diffraction varies significantly in different directions of reciprocal space. This can be corrected by anisotropic scaling, which is especially useful at low symmetry (Rupp, 2009). XPREP is able to scale data anisotropically by making $< |E|^2 >$ in a direction in reciprocal space as similar to 1 as possible. This is often applied to facilitate experimental or MR phasing (McCoy *et al.*, 2007). Macromolecular refinement programs are also able to do "anisotropic scaling" for a model. Here, an anisotropic temperature factor is applied to the entire asymmetric unit. Unfortunately, the parameters of anisotropic scaling are therefore dependent on any TLS (see 5.2.6 on page 63) or other anisotropic displacement parameters. This anisotropy correction is often applied together with bulk solvent correction.

2 Human RNase T2: The MR multi-solution approach

2.1 Introduction

The program SHELXE (Sheldrick, 2002) was designed for experimental phasing of macromolecules and map improvement by density modification. The current beta-test version iterates between density modification and poly-Ala trace generation (Sheldrick, 2010). But the starting phases do not necessarily need to originate from anomalous scattering. An MR solution representing a rather small percentage of the total scattering power can be a sufficient starting point for density modification and main-chain tracing in the new SHELXE, given native data to good resolution. The data from human RNase T2, a protein related both to cancer and brain defects in children (Henneke *et al.*, 2009), was initially intended for S-SAD, but due to weak anomalous scattering and low symmetry, MR was chosen instead as phasing method. Even with potentially good models (sequence homology up to 33%) available, the structure could not immediately be solved. The new version of SHELXE was employed to improve the phases on a number of potential solutions from PHASER (McCoy *et al.*, 2007), which lead to successful solution of the structure.

2.2 Biological background

2.2.1 RNase T2 family

Ribonucleases (RNases) are ubiquitous enzymes that cleave the phosphodiester bond in the ribose-phosphate backbone in RNAs by hydrolysis. They are divided into three main families: A, T1 and T2 (Yoshida, 2001; Deshpande & Shankar, 2002; Raines, 1998). The T2 family consists of acidic endoribonucleases which cleave single stranded RNA, but have no sequence specifity. The catalytic optimum is in the range between pH 4 and 5, with no metal involved into the catalytic activity (Deshpande & Shankar, 2002). They have a typical α/β core structure (Luhtala & Parker, 2010), with the beta sheet consisting of 4–8 strands and the helices forming the exterior of the protein's tertiary structure. A variety of functions has been found for the members of this family: RNA scavenging, RNA degradation, modulation of host immune response as well as cytotoxic functions have been shown to exist. Some of these seem not even to be related to the RNase activity: For example, the plant storage protein CalsepRRP adopts the typical T2 RNase fold while being completely devoid of RNase activity (Rabijns *et al.*, 2002). Most members are glycoproteins (Dieckmann, 2009). The number of disulphide bridges varies between taxonomic kingdoms. Fungal T2 RNases have ten, bacterial ones six; plant and animal T2 RNases have eight cysteine residues (Irie, 1999). Two disulphide bridges are conserved in all members of the family and therefore believed to sustain the active conformation (Deshpande & Shankar, 2002).

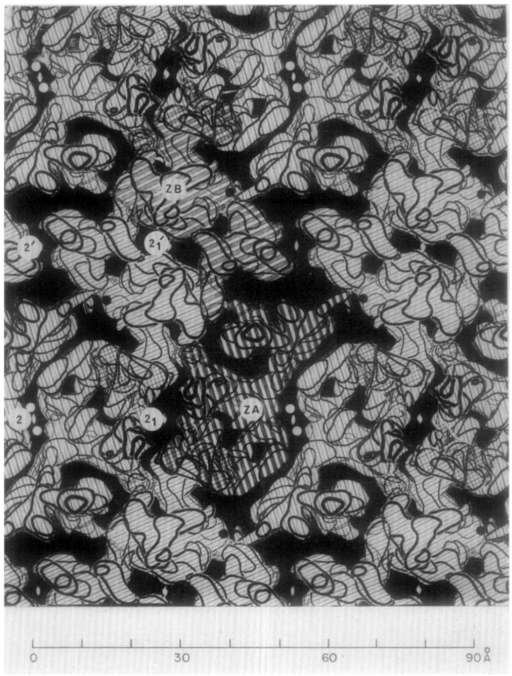

Figure 2.1: Crystal packing figure by Mitsui & Wyckoff (1975). Ribonuclease S was among the first macromolecular crystal structures determined.

T2 RNases cleave RNA internally. In the postulated reaction mechanism, a 2‘,3‘ cyclic phosphate intermediate is formed in transphosphorylation. This intermediate is only released by RNase LE and RNase R. In all other known T2 RNases, hydrolysis occurs and mononucleotides are formed (Deshpande & Shankar, 2002). The reaction is promoted by several histidine residues, which are found in the conserved motives CAS I and CAS II (Luhtala & Parker, 2010). A more detailed account on the reaction mechanism is given in section 2.4.7.

2.2.2 Human RNase T2

Human RNase T2 is the only known human member of this family. The protein is expressed in the brain and in other human tissues as well. Campomenosi *et al.* (2006) postulate a dual lysosomal and secretory role. As human ovarian cancer cells show a decrease in RNase T2 expression (Liu *et al.*, 2002; Acquati *et al.*, 2001), Research implies the full-length-enzyme might have anti-tumoural activity (Acquati *et al.*, 2005; Smirnoff *et al.*, 2006). The tumour suppression might be independent from catalysis, as an enzymatically inactive mutant was shown to suppress tumourigenesis. How this mechanism works remains unclear; it is also assumed that the protein is processed on the way to the lysosome and that only the secreted protein has full length (Campomenosi *et al.*, 2006). Inherited human RNase T2 deficiency possibly causes defects in brain development and angiogenesis as well as leucoencephalopathy. Approximately 40% mass percent of the native protein are heterogeneous glycosylation (Henneke *et al.*, 2009).

2.3 Materials & methods

2.3.1 Preparation

The protein samples for crystallization were provided by R. Krätzner, R. Steinfeld and M. Ziegenbein (Department of Paediatrics II, Georg-August University Göttingen); a short account of the final, improved preparation is given for completeness: Human RNAse T2 was expressed and secreted in HEK 293 cell lines with 1 µg/mL kifunensine to inhibit α-mannosidase. This results in hypermannosylation of the glycosylation chains. The cell supernatant was frozen for later purification. After thawing, the supernatant was directly loaded on a HisTrap HP Ni affinity column (volume 5 mL, flow rate 1 mL/min), and could be eluted as a single peak (buffer A: 20 mM sodium phosphate, 0.5 M NaCl; buffer B: 20 mM sodium phosphate, 0.5 M NaCl, 0.5 M imidazol). The peak fractions were pooled and concentrated using a Millipore Amicon Ultra concentrator. The 88 µL protein solution was digested with 2 µL EndoH (1000 u/mL) in 10 µL NEB G5 buffer for 4 hours at 37°C, leaving N-acetyl glucosaminic residues at the N-glycosylation sites. To lower the content of glycosylated protein in the sample further, a GE Healthcare 1 mL ConA column (flow rate 0.1 mL/min, buffer A: 20 mM TRIS, 0.5 M NaCl, 1 mM $MnCl_2$, 1 mM $CaCl_2$, buffer B: 0,5 M methyl-α-D-glucopyranoside, 20 mM TRIS, 0.5 M NaCl) was used. Again, the protein was concentrated to a volume of 700 µl. Finally, the protein was gelfiltrated using a SuperDex 75 column (flow rate 0.4 mL/min, buffer: 200 mM acetate, 50 mM NaCl, injection volume 100 µl) and could be eluted as a single peak. It was concentrated as before and rebuffered in 20 mM acetate pH 6.0 and 20 mM NaCl. The final protein concentration of 9.2 mg/mL was calculated from a theoretical extinction coefficient, 62045 $M^{-1}cm^{-1}$, based on the protein sequence and the absorption thickness measured using an Eppendorf BigPhotometerPlus photometer.

2.3.2 Crystallization

Crystallization conditions were screened employing both manual pipetting and robot-aided pipetting. For the latter, a well of 100 µL and a drop of 0.1 µL was used on 96-well sitting-drop Greiner plates. The wells were pipetted by a TECAN Genesis RSP 150. The drop was pipetted and mixed with a TTP Labtech Mosquito and consisted of 1:1 protein solution and reservoir. Hanging drop crystallization experiments were set up with Hampton VDXm pre-greased plates (0.6 mL reservoir) and MD CrystalClene cover slips holding a drop of a 1:1 reservoir/protein solution (2 µL) by manual pipetting. All crystals were mounted on MiTeGen MicroMounts and flash cooled by plunging into liquid nitrogen.

Commercial crystallization screens (Hampton Index Screen I+II, Hampton Crystal Screen I+II, Emerald BioSystems Wizard Screen I+II, Qiagen JCSG+) as well as a custom screen consisting of different PEG/buffer mixtures were pipetted by robot for an initial screen on the crude glycosylated protein. Thin needle-shaped crystals (20 µm x 3 µm x 3 µm) formed after four weeks, but could not be mounted on the diffractometer, as they were unstable. The conditions could neither be reproduced nor scaled up. After gel filtration was introduced as the last purification step, the crystallization became reproducible, but the needles did not diffract sufficiently due to small size and high mosaicity.

For the EndoH-digested, deglycosylated protein, new crystallization conditions were found

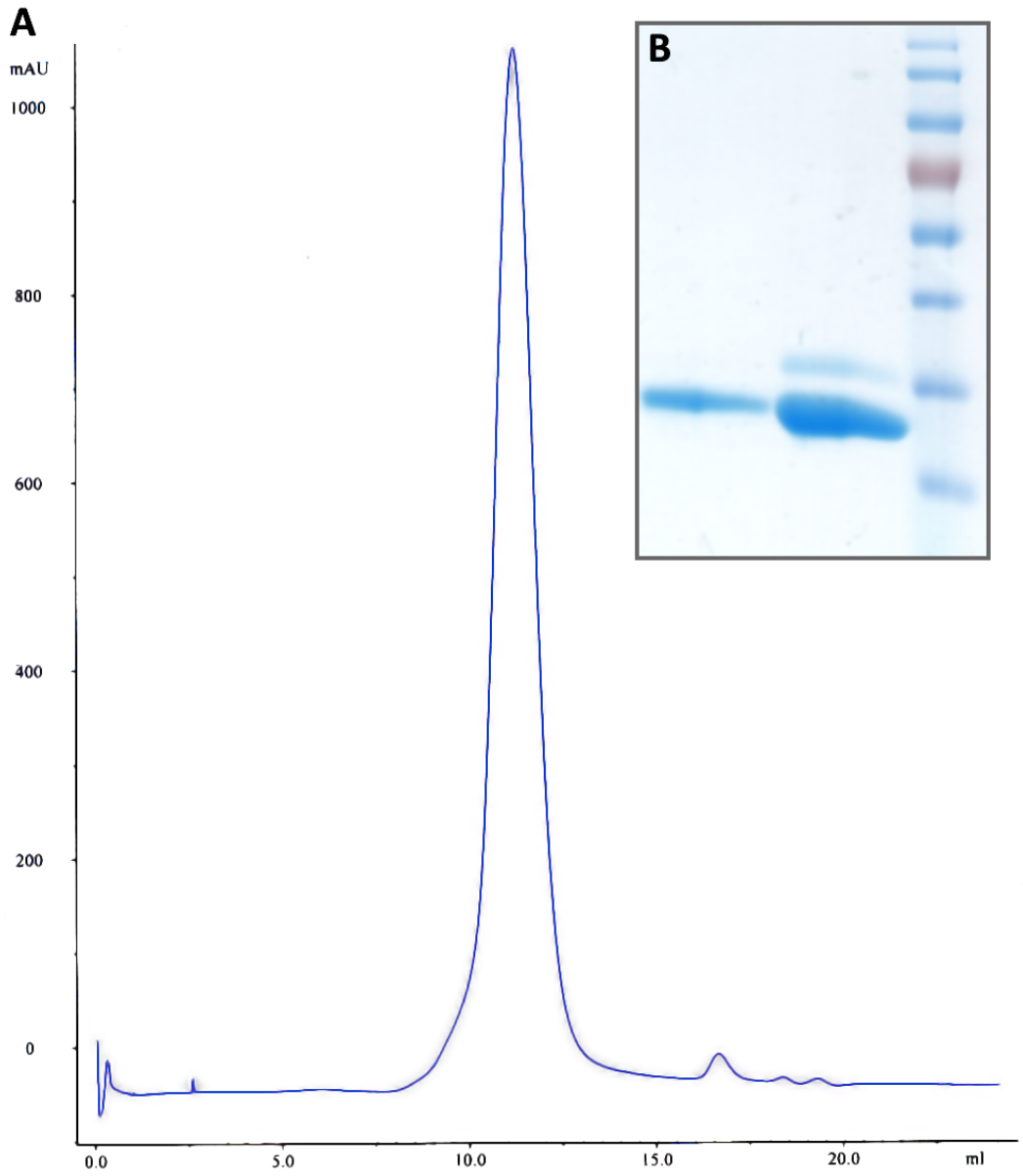

Figure 2.2: **A.** UV-VIS detection in an analytical gel filtration using a SuperDex 75 column (flow rate 0.4 mL/min, buffer as described). **B.** SDS-PAGE of gel-filtrated human RNase T2. (Marker bands refer to specific molecular weight in kDa: 26, 34, 43, 55, 72, 95 and 120)

through extensive screening. An initial monoclinic cell could be determined from crystal plates grown by hanging drop method (reservoir 0.2 M NH_4NO_3 and 20% PEG 3350, drop 1 µL protein solution and 1 µL reservoir, macro-seeded with needles from similar conditions). Approximately 180 needles and plates were screened for diffraction, but no single crystal was found. After two months, block-shaped crystals grew from microcrystals that had formed in a condition of 0.2 M NH_4NO_3 (p.a., Merck), 15% PEG 3350 (Hampton Research) and bi-distilled water. The crystals could not be separated from the viscous precipitate that had formed around it. For cryo protection, the crystals were soaked in a mixture of glucose (99% purity, Merck)/reservoir solution 1:2 (w/v). The crystals had an approximate diameter of 50 µm. Four monoclinic crystals were obtained for X-ray measurement, of which two were single crystals and showed sufficient diffraction.

2.3.3 Data collection and integration

Diffraction screening was carried out at 100K with an in-house source consisting of a Bruker Cu-Kα rotating anode equipped with an INCOATEC multilayer optics, a three circle goniometer and a SMART 6000 CCD detector. Data sets were measured at BESSY MX 14.1 beam line with a Rayonix MX-225 3x3 CCD detector. Data collection statistics are summarized in tables 2.2. The data were integrated with XDS (Kabsch, 2010), converted with XDS2SAD for absorption correction with SADABS (Sheldrick, 2009). XPREP (Sheldrick, 2011) was used for merging, cell determination and data quality analysis unless noted otherwise.

2.3.4 MR models

Four structures (with the highest sequence homology) were chosen from the PDB:

PDB	RNase name	sequence homology	NCBI Blast score
1DIX	RNase Le	34%	113
1IYB	RNase Nw	32%	108
1VCZ	RNase Nt	30%	91
3D3Z	actibind	34%	85

Two additional models were generated with the default options at the SWISS-MODEL homology modelling server (Arnold *et al.*, 2006; Guex *et al.*, 2009; Kiefer *et al.*, 2009). All models were used without ligands and water. To generate a greater variety, the structures were manually trimmed: Sections with high B factors, loops, side-chains and combinations thereof were removed. Additionally, models containing only the consensus core structure and helices were generated. In total, 43 different search fragments were used.

2.3.5 Multi-solution approach with PHASER and SHELXE

All native data sets were merged into one file, which was used for all MR and SHELXE runs. A variety of MR solutions was generated with PHASER (version 2.1.4, McCoy *et al.*, 2007), employing the aforementioned different models, as well as different resolution cut-offs. The positioned models from these solutions gave starting phase information for the iterative density

modification and poly-Ala tracing in SHELXE. The resolution cut-off for initial phases from the input model was set to 1.8, 2.0, 2.2., 2.5, 3.0, 3.5 and 4.0 Å, respectively, and two solvent contents were used: 45%, the default value in SHELXE, and 30%, the approximate value to be expected from the sequence, the unit cell volume and with 1 molecule assumed in the ASU.

2.3.6 General test of the multi-solution approach

A data set of concanavalin A (Hardman & Ainsworth, 1972) measured in our lab has been used. The structures used as models are given in the appendix on page 87 and were chosen by sequential alignment and scoring with PSI-BLAST (Altschul *et al.*, 1997). Either fragments of this PDB entries or the full protein was used for a number of PHASER runs. Only solutions with a TFZ lower than 8.5 were used for SHELXE. The solvent content was set to the default value (45%) and no helix search was employed as concanavalin A doesn't contain helices. The resolution cut-off for starting phases was varied (1.6, 2.0, 2.5 and 3.0 Å). The results are tabulated in the appendix on page 87.

2.3.7 Final solution and trace optimization

The structure solution that was later used for refinement has was obtained from RNase Le (Tanaka *et al.*, 2000) as search model (without any trimming). The correct PHASER solution was identified and improved by density modification in SHELXE and subsequent poly-Ala tracing (CC against native data = 29.30, average chain length = 27.2). The resulting backbone trace was optimized by recycling it as input SHELXE.

Table 2.1: Trace optimization by varying the solvent content (command: **`shelxe -a -q -e1 -m30 -l3 -s[fractional]`**).

solvent	CC vs. native data	average chain length	residues total
25%	28.42%	21.0	168
30%	30.21%	25.4	178
40%	26.39%	28.2	169

From these solutions, the second was chosen as the best trace. A SHELXE *free lunch* map extended to 1.0 Å resolution was used for the first refinement model.

2.3.8 Refinement and structure validation

The structure was refined with REFMAC (Murshudov *et al.*, 1997), but showed overall too long bond lengths. The structure was checked with the "anomalous bond length test" at the WHATIF web service (Rodriguez *et al.*, 1998; Vriend, 1990). A new cell obtained by iteratively refining and checking again with this test. Integration was repeated with the corrected cell and the model adapted using the resulting data. After check with the TLS-MD server (Painter & Merritt, 2006), TLS refinement (Winn *et al.*, 2001) was applied, with only one domain consisting of the

whole protein and the two N-acetyl glucosamine residues. The structure was validated with MOLPROBITY and the weighting scheme optimized using different weights and 100 refinement cycles to ensure convergence. No residues were observed in the generous and in the disallowed regions of the Ramachandran plot (Ramachandran & Sasisekharan, 1968). The structure was subjected to the SSM web-service (Krissinel & Henrick, 2004). No significant intermolecular contact surface could be found, as the molecules seem to be biological monomers.

2.4 Results

2.4.1 Crystallization

Crystallization of T2 RNases often proves difficult, as crystals grow as thin plates or needles unsuitable for X-ray diffraction data collection (Deshpande & Shankar, 2002). This holds true for human RNase T2, and only after several thousand conditions, we were able to achieve crystals suitable for data collection. Deglycosylated protein as well as gel filtration as the final purification step were necessary prerequisites. The protein solution did not tolerate buffer or heavy metal ions. Originally, we aimed for crystallization of the glycosylated protein (Mesters & Hilgenfeld, 2007), as the sugar moieties might be vital to function and specifity of human RNase T2. Despite our efforts, only deglycosylation lowered the surface entropy enough to achieve crystals suitable for measurement.

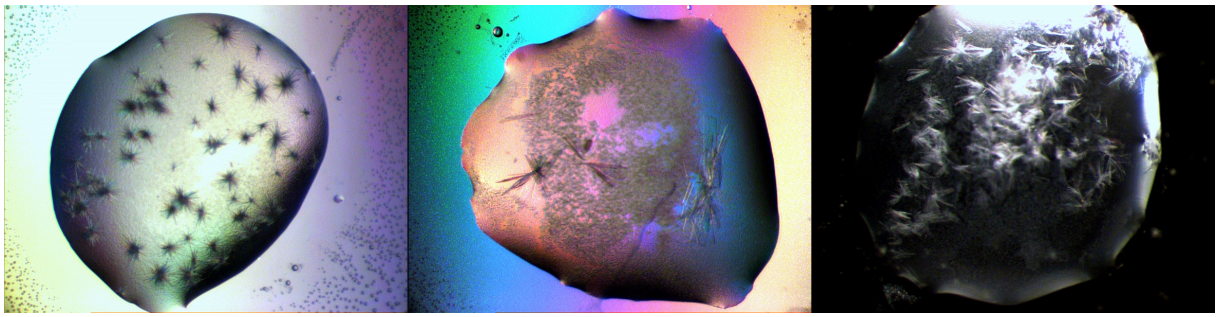

Figure 2.3: Typical crystals from human RNase T2; the drops measure approximately 2 mm across. No single crystals could be retrieved from such conditions.

2.4.2 Data collection and integration

Three native data sets and two long-wavelength data sets were collected. For the integration of native data, initially, the wrong wavelength had been used, resulting in a cell with axes 7% too long axes. Tables 2.2 and 2.3 refer to data re-integrated with the correct cell. Due to the low symmetry space group and low multiplicity, the anomalous signal was weak. It was not used for S-SAD, but for validation, as discussed in chapter 4 on page 51.

Table 2.2: Summary of data collection statistics for native data. Values in parentheses refer to outer resolution shell.

	rnase3_ds	**rnase3_ds2**	**rnase3_ds4**
source	BESSY MX 14.1		
unit cell dimensions	a = 31.32 Å , b = 68.09 Å, c = 47.98 Å, β= 90.83°		
space group	$P2_1$		
wavelength (Å)	0.9184	0.9184	0.9810
oscillation range	95°	103.5°	237°
resolution range (Å)	27.56-1.73 (1.83-1.73)	28.31-1.74 (1.84 -1.74)	28.31-1.59(1.69-1.59)
no. of observations	40334 (5303)	43694 (6281)	75521 (11987)
unique*	20100 (2714)	20030 (2926)	26238 (4246)
multiplicity*	1.96 (1.73)	2.13 (1.94)	2.80 (2.55)
completeness* (%)	97.5 (88.6)	97.7 (90.5)	97.2 (90.3)
mean I/σ(I)	17.00 (4.11)	20.47 (6.48)	16.74 (5.04)
R_{int} (%)**	3.06 (19.54)	2.63 (13.59)	3.07 (19.10)
R_{rim} (%)**	4.15 (26.80)	3.45 (17.83)	3.75 (23.38)
R_{pim} (%)**	2.79 (18.25)	2.21 (11.44)	2.11 (13.31)

* Friedel pairs merged.
** As defined in the appendix on page on page 85.

Table 2.3: Summary of the collection statistics for anomalous data. Values in parentheses refer to outer resolution shell.

	rnase3_ds3	**rnase2_ds2**
source	BESSY MX 14.1	
unit cell dimensions	a = 31.33 Å, b = 68.15 Å, c = 47.99 Å, β = 90.83°	a = 31.56 Å, b = 69.44 Å, c = 48.37 Å, β = 90.54°
space group	$P2_1$	$P2_1$
wavelength (Å)	1.950	1.950
oscillation range	95°	103.5°
resolution range (Å)	47.90–2.23 (2.33–2.23)	47.90–2.43 (2.53–2.43)
no. of observations	68336 (6068)	26598 (14722)
unique*	18256 (1919)	13434 (1017)
multiplicity*	3.58 (2.54)	1.81 (1.17)
completeness* (%)	95.5 (80.2)	91.3 (61.4)
mean I/σ(I)	13.66 (1.98)	9.25 (1.63)
R_{int} (%)	8.00 (49.44)	8.72 (44.12)
R_{rim} (%)	9.39 (58.24)	11.59 (58.60)
R_{pim} (%)	3.47 (22.98)	6.00 (31.65)
R_{anom} (%)	8.33 (66.99)	13.13 (81.88)
d''/σ(d'')	0.88 (0.87)	0.89 (0.90)

* Friedel pairs not merged.
** As defined in the appendix on page on page 85.

2.4.3 MR multi-solution approach on human RNase T2

A number of trimmed models were generated from four structures and two homology models from SWISSPROT (Arnold *et al.*, 2006; Guex *et al.*, 2009). 42 runs of PHASER yielded solutions with translation function Z-scores between 2.7 and 5.6, meaning no definitive solution occurred (see section 1.2.1 on page 9). The translation Z-score might have been low due to the monoclinic space group. From each run, the placed model of the solution with the highest LLG was used for initial phases in SHELXE. Every five cycles of density modification (using the *sphere of influence* algorithm) the resulting map was used for automatic poly-Ala tracing. To ensure a relatively big loss of bias from the original structure, 15 such iterations were executed in total. Among the tests, two parameters were varied: The resolution cut-off for the model giving start phases and the solvent content for the density modification. Of a total of 588 SHELXE runs, three yielded a correct phase solution. As shown in Fig. 2.4, the solved trials clearly stand out against the unsolved ones in both average chain length as well as CC against native data.

They resulted from PHASER solutions with relatively low TFZ (see Fig. 2.4). With the lower solvent content (calculated from the sequence and the unit cell volume, 35%) no solution could be found. After refinement of the structure, it became clear that due to missing residues and floppy regions, the disordered region of the crystal is in fact 44%, a percentage near to the default value in SHELXE of 45%. Also, density modification generally works better for structures with a high solvent content, which should be given accurately or slighly overestimated (compare section 3.4.4 on page 43). Concerning the starting phases resolution cut-off, here, only relatively good resolutions (1.8 Å and 2.0 Å) lead to a structure solution.

It became clear why MR did not work well in the first place, and only MR in combination with SHELXE led to a structure solution: Molecular replacement, as it is dependent on Patterson peaks from long intra-molecular inter-atomic vectors, is more susceptible to cell distortion than the SHELXE auto tracing algorithm which places one Ala fragment after the other and can so compensate more easily.

It should be noted that 46% of the structures in the PDB tested for the PDBREPORT (Hooft *et al.*, 1996; Joosten, personal communication) were flagged by the "anomalous bond length test" for having a somehow distorted cell. While this does not mean necessarily a cell so drastically misdetermined as in this case, it gives a good indication that this is not an uncommon error and might be a reason if MR fails. SHELXE can help to overcome this problem and the cell can be checked after structure solution.

2.4.4 A principal try on the multi-solution approach

In the case of a distorted or misdetermined cell, the multi-solution approach using SHELXE can overcome the problems in MR. But whether it can also improve the phasing in case of a correct cell remained unclear. We used one of the structures from the **`REST`** test library (see Table 5.3.1 on page 66) for a general test. 16 PHASER solutions with a TFZ ranging from 4.1 to 8.8 were chosen and read into SHELXE for 30 iterations of auto tracing and density modification. The higher number of iterations ensures a better resolution of successful trials from unsuccessful ones. No helix search was employed here, as the chosen structure of concanavalin A does

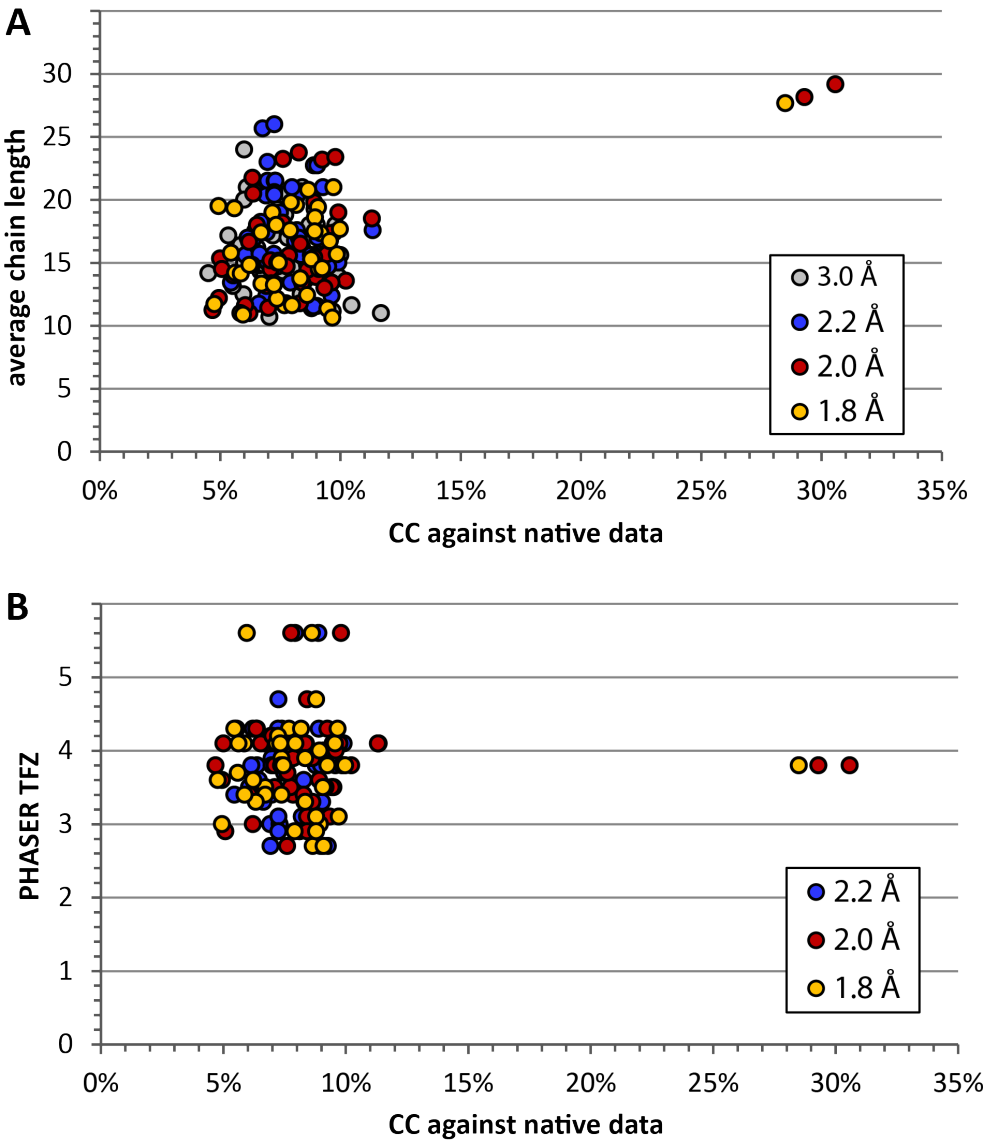

Figure 2.4: For human RNase T2: **A.** Scatter plot of SHELXE runs for selected **PDB** resolution cut-offs. The three solutions clearly stand out against the not sucessfully traced trials. **B.** Scatter plot against PHASER TFZ. The solutions do not result from the PHASER runs with the highest TFZ, but this might be an effect of the distorted cell. This confirms the program's author's assumption that structures with a CC against native data over 25% are clearly solved. For the average chain length, however, it was stated that a value over 10 hints to a correct solution (Sheldrick, personal communication). Almost all trials ended with the average chain length over 10, but the successful solutions have over 25, at least for this case.

not contain helices. Four runs from three different PHASER solutions yielded correct phase solutions. (As with human RNase T2, the potential solutions were compared in COOT (Emsley & Cowtan, 2004; Emsley *et al.*, 2010) with the refined structure.) The runs are shown as scatter plot in Fig. 2.6. Here as well, the plot of CC (native) against the average chain length proves the criterion of "average chain length higher than 10 hints at a solution" is proven incorrect. CC against native data alone is a better indicator: All solutions with a CC > 25% are correct and all solutions < 25% are incorrect.

We employed only regular PCs and MR solutions devised in the usual way from PHASER, as opposed to ARCIMBOLDO, where helix fragments are used for an *ab-initio* approach on a CONDOR-run computer grid with much higher performance.

In the plot against the PHASER translation function Z-score, the two highest ranking PHASER solutions also yield three solutions. The fourth solution, however, is from a PHASER run with only a TFZ of 4.8 and a LLG of 35. Here, MR was not successful; nonetheless, if used as input to SHELXE, a correct phase solution can be gained. This resembles the method of *Patterson seeding*, as used in small molecule crystallography and hence could be called "MR seeding", for the used fragments are not solutions by themselves, but not completely random either.

If the CC against native data is plotted against trace iteration (Fig. 2.5), two interesting features become apparent: The start values already indicate a potentially successful run. And the CC varies until it starts to increase rapidly and then varies within a higher value range. Such progression is commonly seen in small molecule direct methods, such as charge flipping (compare *e.g.* Oszlányi & Süto 2004). Non-successful traces vary, and even decrease in their CC value.

With this, we prove the general principle, but more tests on a variety of structures are needed. We aim for a routine method to combine phase information from MR and density modification.

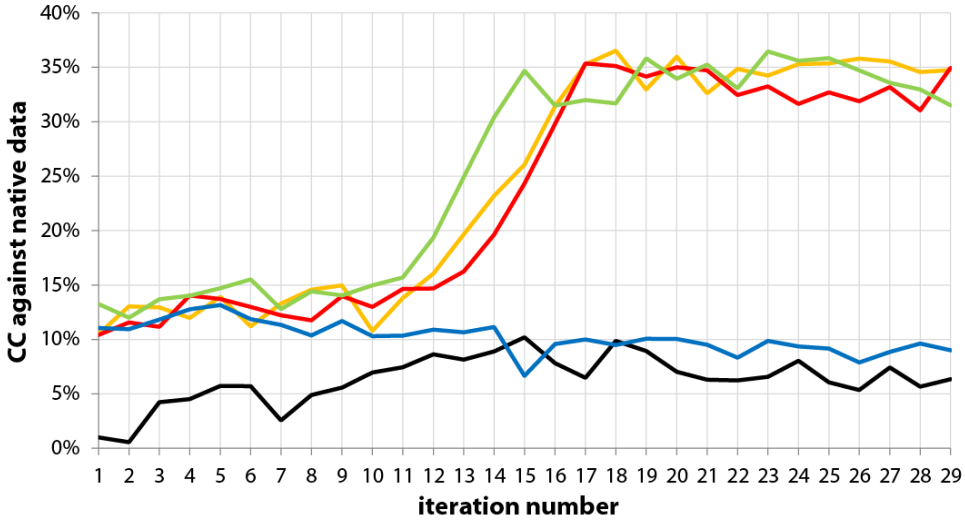

Figure 2.5: Progression of CC values along trace iterations for concanavalin A in five different SHELXE runs. The successful traces start at high values and after a few iterations progress steadily into a higher range. The blue trace starts at a relatively high value, but does not lock. Low start values may indicate that a successful trace is unlikely.

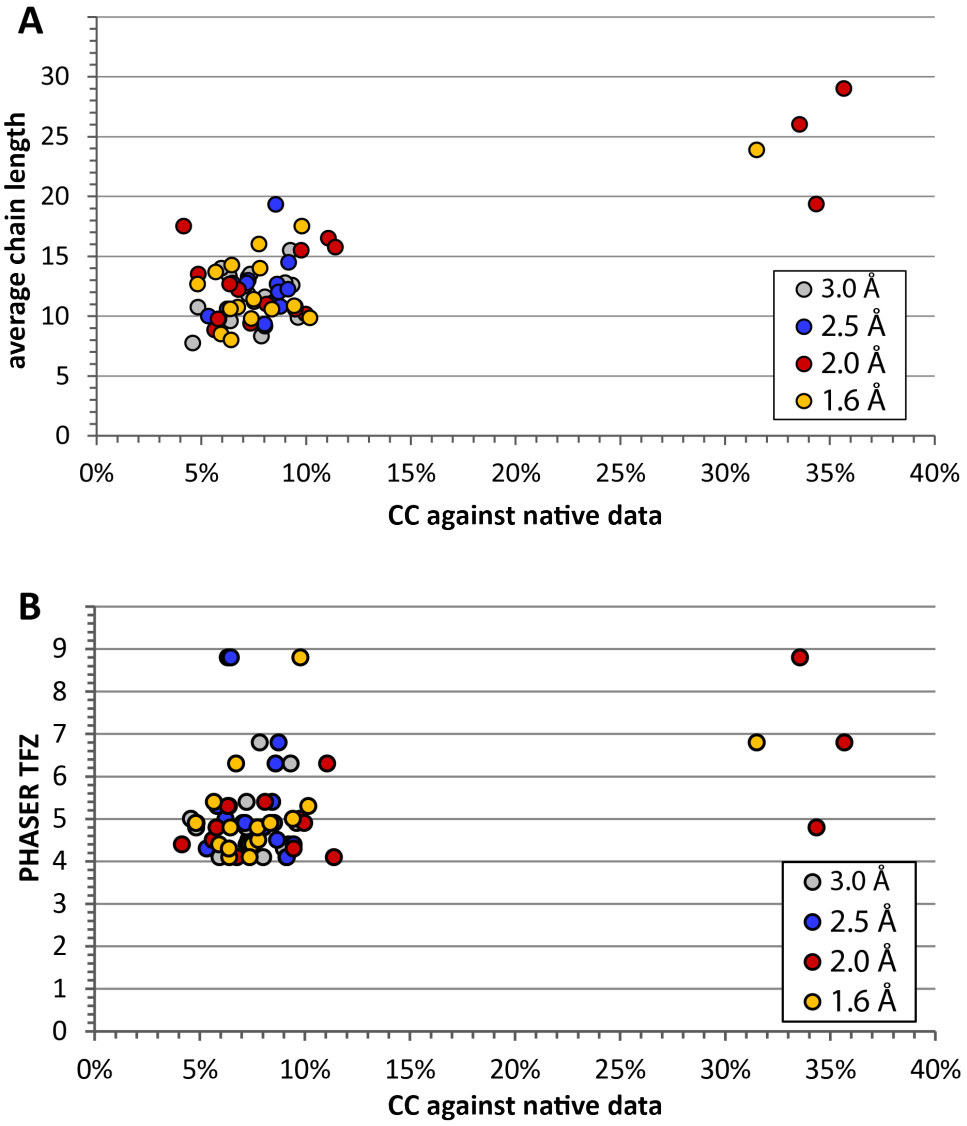

Figure 2.6: For concanavalin A: **A.** Average chain length against CC (native data) for selected resolution cut-offs. **B.** PHASER TFZ against CC (native data).

2.4.5 SHELXE trace optimization and refinement

One of the three successful solutions was again subjected again to a number of SHELXE runs for poly-Ala tracing, but with more solvent content variation and extended tracing options.

The best trace was obtained with a solvent content of 30%. In the **PHS** map, which had been extended to 1.0 Å by *free lunch* algorithm, the initial model for refinement with 178 full residues was built from the poly-Ala backbone using COOT.

The structure was refined with COOT and REFMAC. TLS refinement was applied and both checks with WHATIF as well as with MOLPROBITY lead to a significant improvement of the model. At two glycosylation sites, N-acetyl glucosamine residues could be found.

The weighting scheme was tested against the negative log likelihood gain minimum, R_{free} and against the MOLPROBITY score. The final R values as well as other quality indicators and statistical values are given in Table 2.4. The structure shows a typical T2 RNase fold, with four disulphide bridges (including cysteine residues 48/55, 75/121, 184/241 and 202/213) and an α/β core motif, as shown in Fig. 2.7.

Table 2.4: Refinement statistics.

refinement statistics	
resolution range	27.56–1.59
working set reflection number	25064
working set completeness (%)	98.0
N of reflections in test set	1347
solvent content (%)	43.99
no. of protein atoms	1642
no. of water molecules	182
protein molecules per ASU	1
R (%)	15.38
R_{free} (%)	19.03
average B factors (Å^2)	
overall	18.89
protein atoms	18.27
water molecules	21.93
r.m.s.d. from ideal geometry	
bond lengths (Å)	0.021
bond angles (°)	1.057
Ramachandran plot, residues*	
in favoured regions	97.37
in allowed regions	2.63

*Calculated with MOLPROBITY

2.4.6 Comparison with similar proteins

The structure was aligned with the entries of the PDB with the SSM tool (Krissinel & Henrick, 2004). The best results are shown in Fig. 2.8. While the core fold is strictly conserved, the outer loops, especially residues 185–194, show differences between the structures. In human RNase T2, this loop could only be partially modelled and has high B factors proving its flexibility.

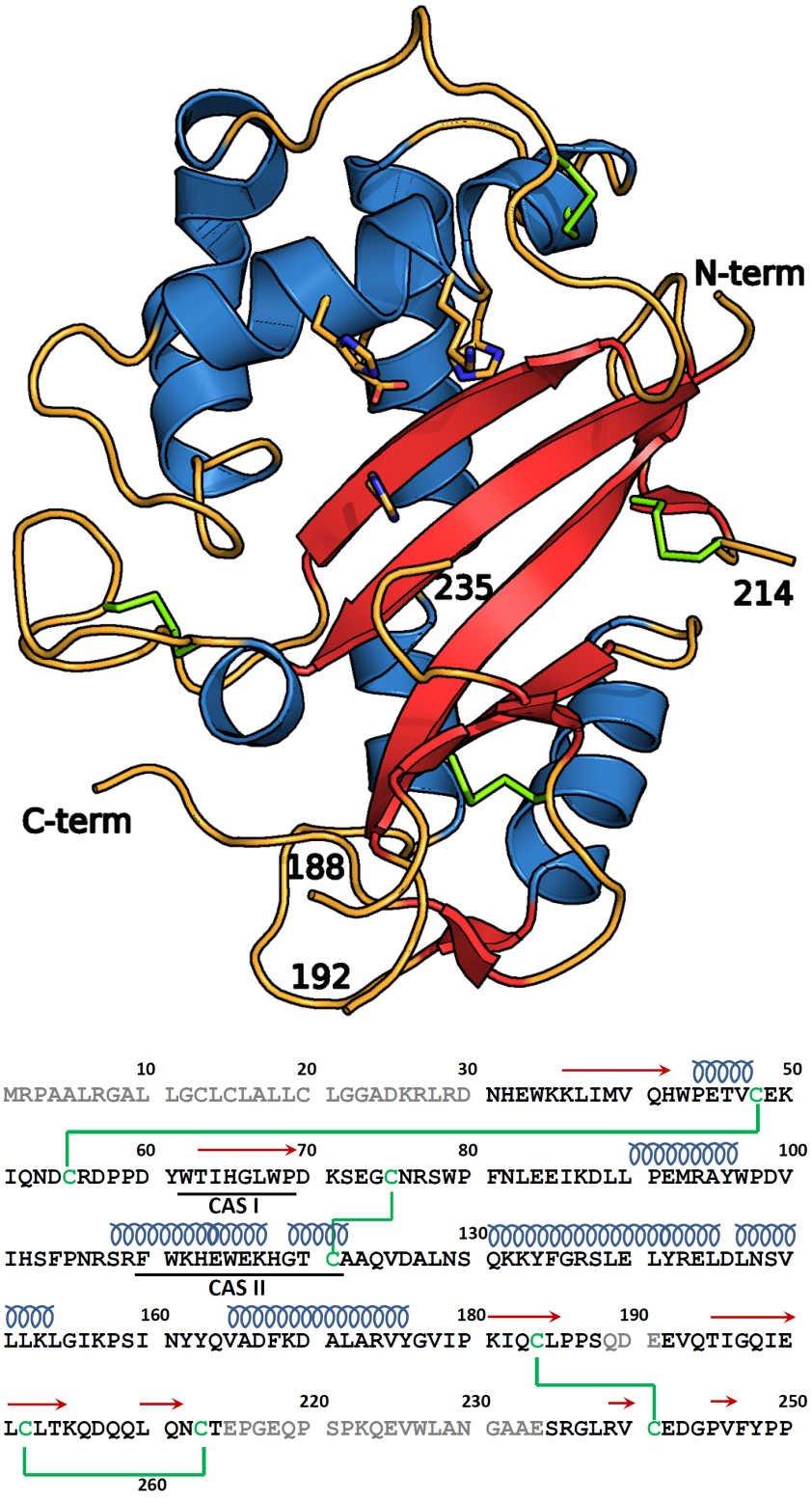

Figure 2.7: Cartoon representation of the final model. The active site residues are shown as sticks, disulphide bridges in green. The colors reference the secondary structure to the sequence with active site motifs CAS I and CAS II below. Disulphide bridges are marked in green. Residues from the cDNA sequence that could not be found in the density are grey.

Name	PDB	residues	SSM Q score	sequence identity	r.m.s.d.
RNase MC1*	1UCG	190	0.64	28%	1.62
RNase Le	1DIX	208	0.64	33%	1.50
RNase NW	1IYB	208	0.63	31%	1.66

*The mutant N71T was chosen as the structure shows a slightly lower r.m.s.d. with our protein then the wild type. All r.m.s.d. values and the given sequence identity are in comparison with RNase T2 (in grey), but only for the sequence part used by SSM tool.

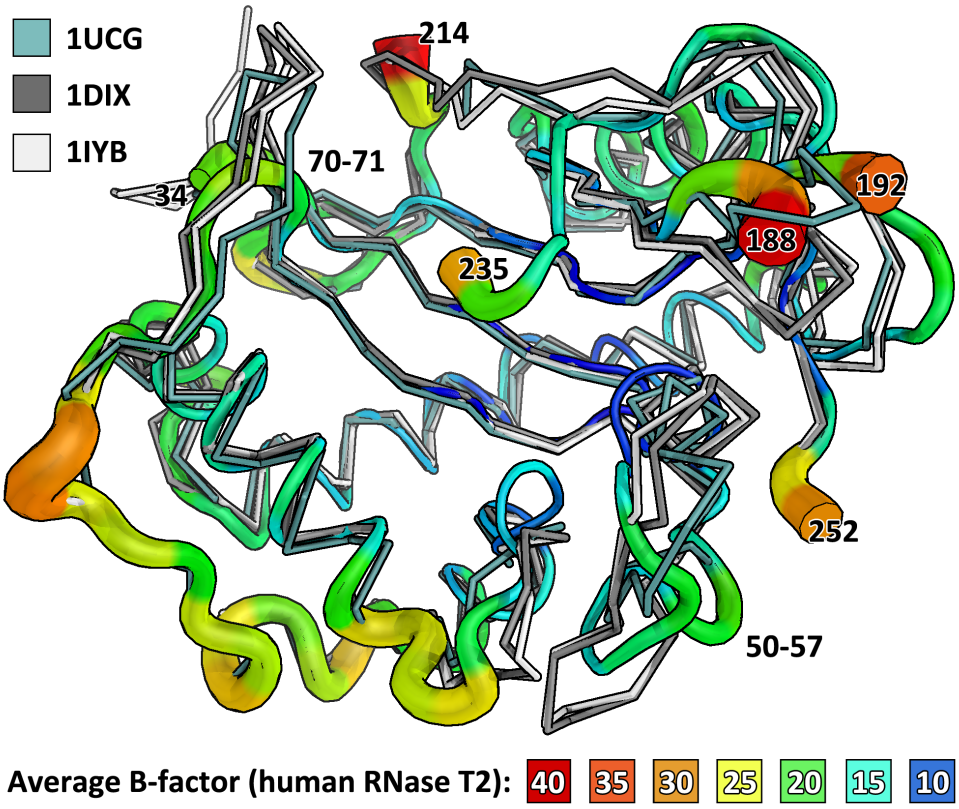

Figure 2.8: Overlay of the B factor putty representation of human RNase T2 and three other members of the its family: RNase MC1 (1UCG) in pale teal, RNase Le (1DIX) in grey and RNase NW (1IYB) in light grey. Note that the other structures have a prolonged loop between residues 70-71. Also, the disordered loop 214-235 seems not to have an equivalent in the other structures, and possibly poses an insertion. Loop 50-57 is bending inwards as compared to the other T2 RNases. Apart from these differences in the outer regions, the core fold is highly conserved and rigid, as can be seen from its low average B factor.

2.4.7 Overall structure and reaction mechanism

Two common motifs are to be found in T2 RNases, CAS I and CAS II. Irie et al. (1997; 1999) proposed the mechanism for an acid-base reaction in RNase Le. As the structure of the active site is well conserved in T2 RNases (see Fig. 2.9), the same reaction mechanism can be assumed for human RNase T2 (see Fig. 2.10). The initial cleavage and cyclization is promoted by His 65, His 113 and His 118. Lys 117 and Glu 114 stabilize the the intermediate five-membered ring. Hydrolysis occurs in the second step. The alternative conformation of Lys 117 in the human RNase T2 structure might be a result of the high side-chain flexibility.

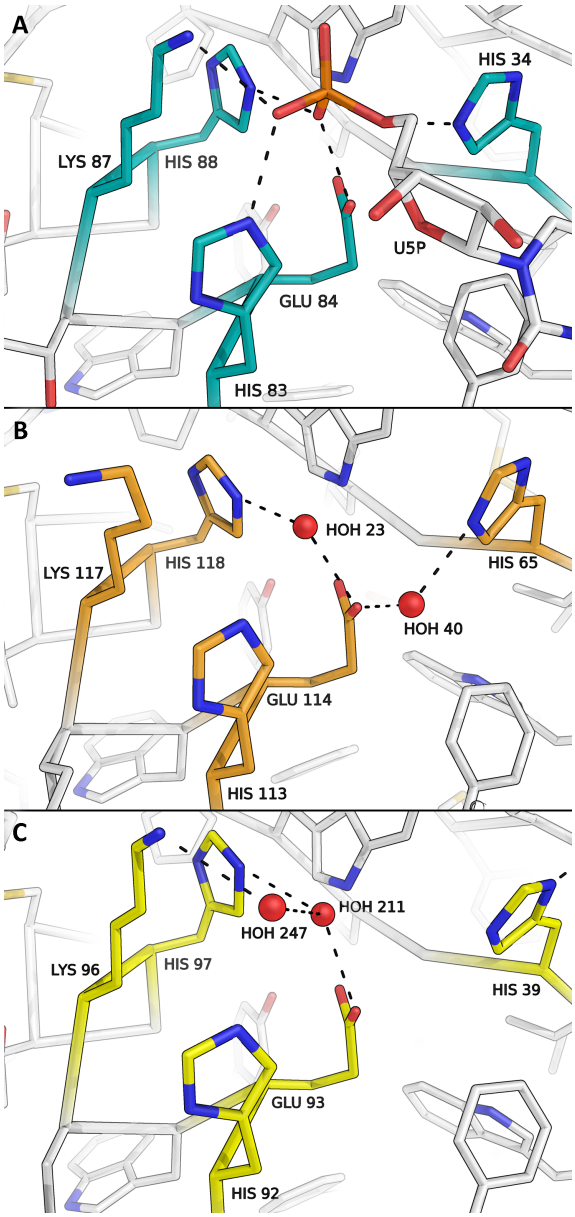

Figure 2.9: T2 RNAse active sites: **A.** RNase MC 1 with bound 5′ UMP (PDB 1UCD). **B.** Human RNase T2. **C.** RNase Le (PDB 1DIX).

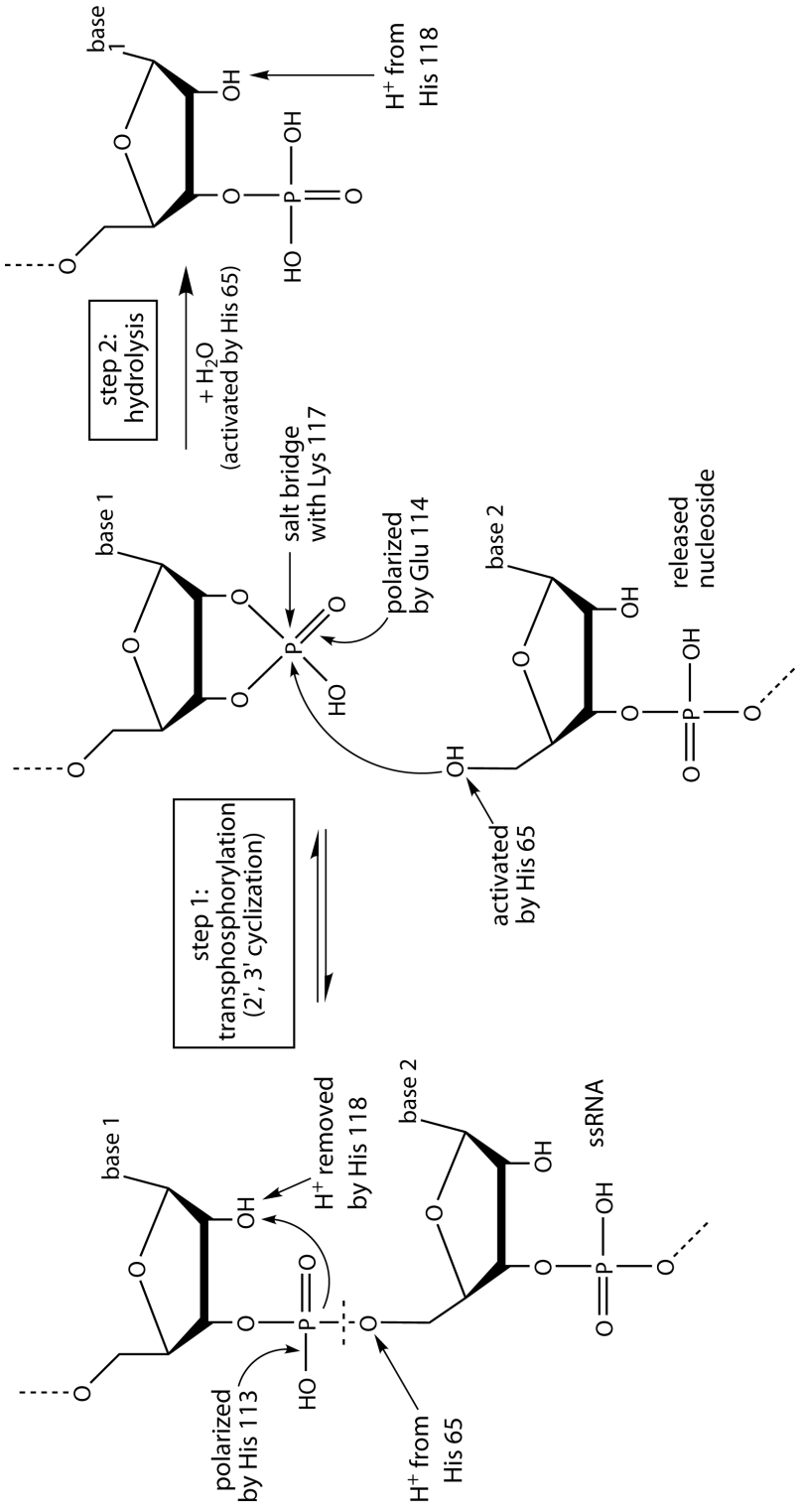

Figure 2.10: Assumed reaction mechanism for human RNase T2.

2.4.8 Missing residues and mass spectrometry

Several parts of the protein were not visible in the electron density. Mass spectrometry gave a mass lower than expected from the sequence and the detected fragments gave no evidence for the first 27 N-terminal residues. The sequence was confimed by sequencing the cDNA as well as the transcript mRNA of the inserted construct. Therefore, the protein must have been post-translationally modified. The Signal Peptide Repository (Gasteiger *et al.*, 2003; Boeckmann *et al.*, 2003; Consortium, 2011) lists the first 24 residues of human RNase T2 as a potential signal peptide. Signal peptides are a common feature among secretory proteins: They target the protein of the endoplasmic reticulum and into the secretory pathway. Usually, after the ER membrane is passed, signal peptidase cleaves the signal peptide from the main protein. (Blobel & Dobberstein, 1975; Martoglio & Dobberstein, 1998).

To further clarify this, the protein was sequenced using endoprotease digestion and electrospray ionization mass spectrometry by Henning Urlaub, Uwe Plessmann and He-Hsuan Hsiao (see Fig. 2.11). The sequence without the signal peptide was confirmed exactly. Residues 189–191 and 215–234 were confirmed by mass, but could not be modelled in the density: They were disordered, and belonged to the flexible surface of the protein.

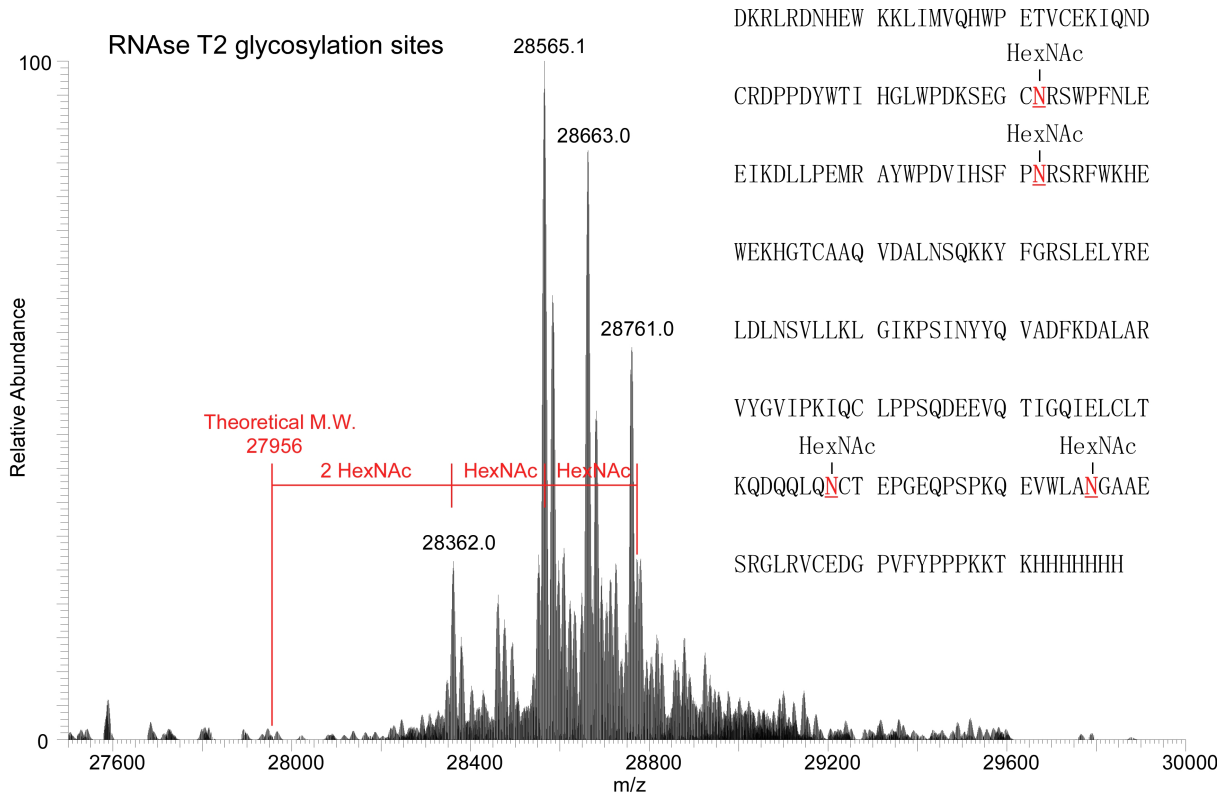

Figure 2.11: Mass spectrum for digested human RNase T2 with the peaks for glycosylated species marked. Figure by He-Hsuan Hsiao.

2.4.9 Glycosylation

Control SDS-PAGE from the EndoH digestion as well as prediction by the NetNGlyc server (Gupta *et al.*, 2002) indicates the existence of three glycosylation sites at Asn 106, Asn 76 and Asn 212. N-acetyl glucosamine residues bound to Asn 76 and 212 left over from the deglycosylation were clearly visible in the density and were modelled giving further evidence for these glycosylation sites. For Asn 106, the density indicated no left-over sugar. The Fig. 2.12 illustrates these three glycosylation sites and their residual electron density at 1.0σ. As the protein was sequenced by mass spectrometry, these three sites were confirmed and a fourth site was found: Asn 230. As this lies in one of the two disordered loops of the protein, no crystallographic account for this site can be given. This glycosylation site could not be found in the wild-type protein, and might occur because of the hypermannosylation in the production of the protein.

2.5 Outlook

Using a SHELXE multi-solution approach, the structure of human RNase T2 could be solved. In cases where MR cannot clearly solve a structure, or the correct solution is not clearly indicated, density modification and auto tracing with SHELXE could provide additional phase information and clearly point to the correct solution. It is a very robust treatment, as shown with RNase T2, where the distorted cell obscured the Patterson search in PHASER (McCoy *et al.*, 2007), but auto tracing succeeded even with the poor MR solutions provided. Also, model bias might be reduced due to additional phase information from SHELXE. This is exploited already in the program ARCIMBOLDO (Rodríguez *et al.*, 2009), and could become a routine procedure for cases where MR cannot clearly solve a structure. With concanavalin A as test structure, a proof-of-principle was given, although more tests are needed to develop a general method.

There might even be cases where the major amount of phase information is derived from repeated it-

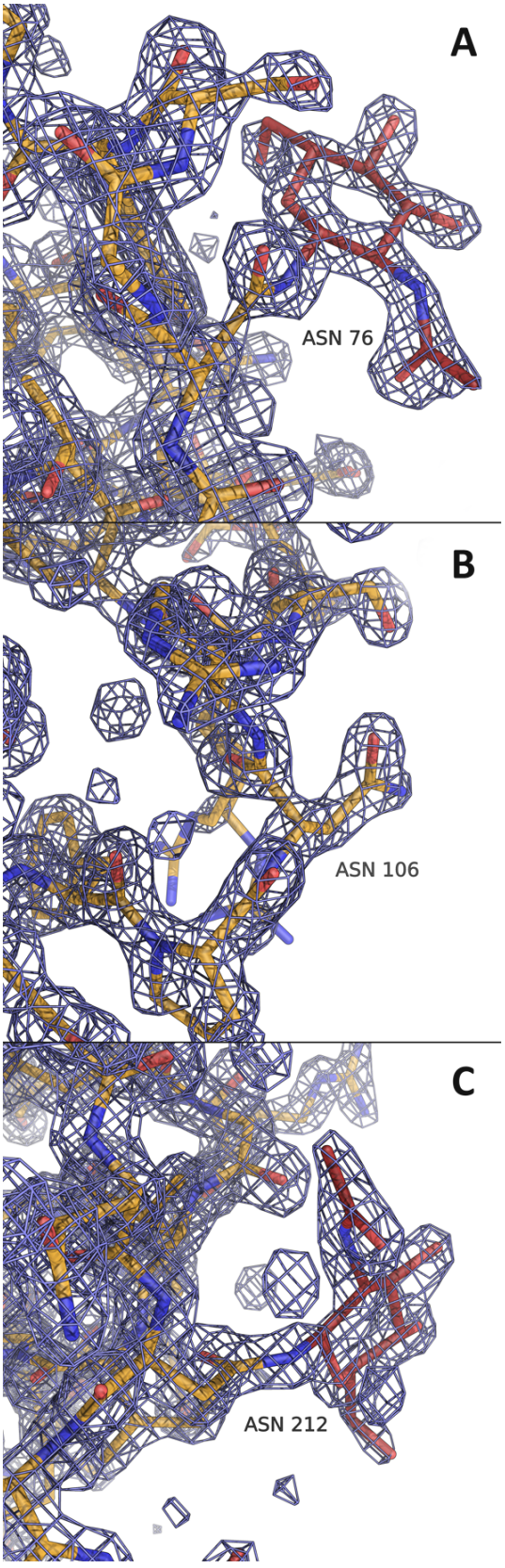

Figure 2.12: Glycosylation sites: **A.** Asn 76 **B.** Asn 106 **C.** Asn 212

erations of SHELXE density modification and auto tracing, and the MR solution just provides somewhat better-than-random starting phases. Such a multi-solution approach is already known from small molecule direct methods, where starting phases are obtained by *Patterson seeding*. By analogy, the MR multi-solution approach could be called "MR seeding".

We could determine the X-ray structure of human RNase T2, and confirmed the existence of a signal peptide in the sequence as well as four glycosylation sites, of which one is in an uncommon motif. The structure has not been completely interpreted yet and will be subject to further research.

3 Hellethionin D: MR-SAD

3.1 Introduction

For small molecules, the phase problem can almost always be solved by direct or Patterson methods. For macromolecules with their less ordered crystals, worse resolution and much bigger structures, still, the phase problem is one of the great challenges in structure solution. We have a great variety of methods at our hands – among them experimental phasing methods and molecular replacement.

While these two "realms" have evolved greatly, only recently their combination has gained wider attention. Today, the software and our knowledge of the phase problem allow us to combine our prior knowledge of solved structures with the phase information from experimental methods. If one of the established methods alone fails, combining phase information from several sources might give the little more phase information needed to lock in to a correct solution.

In this work, we applied MR-SAD (Schuermann & Tanner, 2003) to solve Hellethionin D from *Helleborus purpurascens*. We used the NMR structure of the protein as search model, which was positioned in the unit cell by using a modified version of ARCIMBOLDO. After this, we employed density modification and S-SAD to further improve the phases with SHELXE. The result was a trace of 299 of 318 protein residues in the ASU.

3.2 Biological background

Hellethionin D from *Helleborus purpurascens* (Fig. 3.1) is a typical thionin in length and fold. Thionins are inhibiting and anti-pathogenic peptides of approximately 46 residues length. They feature 3–4 disulphide bridges, a very robust tertiary structure and often a positively charged loop region. Examples of this class are viscotoxins, purothionins and crambin. Thionins have been shown to be toxic in vitro to bacteria, fungi and yeast, and therefore are thought to be part of the pathogen defence in the plant (Milbradt *et al.*, 2003). Agricultural transgenic plants that can express thionins for enhanced resistance against microbes have been patented (Ohashi *et al.*, 2001).

Figure 3.1: *Helleborus purpurascens*.
Image courtesy of Zdeněk Pazdera.

In 2003, the NMR structure of Hellethionin D was solved by NMR (Milbradt *et al.*, 2003). The 20 lowest energy structures are deposited as PDB entry 1NBL.

Significant differences to other γ-thionins are assumed to be related to the unusual charge distribution and the threonine-rich sequence 36-39 of hellethionin (Milbradt *et al.*, 2003): "In fact, the well-defined 3D structure of hellethionin D is very similar to those reported so far for viscotoxins, purothionins, or crambin, although distinct differences could be detected in the C-terminal portion, especially for loop 36–39. These differences may derive from the unusual distribution of charged residues in the C-terminal half of the peptide sequence compared to other thionins and from the uncommon occurrence of four contiguous threonine residues in loop 36–39."

3.3 Materials & methods

3.3.1 Preparation and purification

The purified and lyophilized protein was provided by F. Kerek and co-workers (DoNatur GmbH, Munich). An overview of the preparation is given for completeness: All chemicals were used as supplied in *pro analysi* quality from Merck, if not mentioned otherwise. For extraction at room temperature, 2 kg dried root and rootstock of *Helleborus purpurascens* were coarsely milled, treated with hexane and air-dried. The defatted plant material was extracted with a mixture of water/ethanol/acetic acid (39:10:1). The filtered extracts were pooled and concentrated to a volume of 1.0 L by evaporation (70°C, vacuum), treated for 2 h with 35 g active coal and filtered again. The filtrate was stirred into a tenfold volume of cooled (10°C) acetone and the brownish-grey precipitate was separated by centrifugation (4000 rpm). This was repeated until the supernatant was only yellowish coloured. The final yield of raw product (5.6 g) were dissolved in 200 mL of de-ionized water and passed through an anion-exchange column (Sigma-Aldrich Ambersep-900) to retain anionic impurities. The raw alkaline (pH 11) solution of hellethionin was acidified to pH 3 by treatment with the adequate amount of strong cationic resin Ambelite 120 (Sigma-Aldrich, previously treated with 1 M HCl and washed with water). The filtered solution was lyophilized.

From the lyophilized crude extract a 10 mg/mL solution with 0.08% (v/v) trifluoroacetic acid and 20% (v/v) acetonitrile was prepared. 0.4 mL aliquots were injected on a Macherey-Nagel Dueren Nucleosil 100-7 C8 column (250 mm length / 21 mm diameter, flow rate 3 mL/min). The Bio-Tek Kontron HPLC system consisted of a pump 422, a gradient 425 former, and a UV-Detector 430. A linear gradient of buffer B from 20% to 50% in 30 minutes was applied (buffer A = 0.1% trifluoracetic acid, buffer B = acetonitrile, 0.1% trifluoracetic acid). Collected fractions were: Hellethionin A (14.4 ± 0.3 min), Hellethionins B1–B6 (16.1 ± 0.6 min), Hellethionin C (16.9 ± 0.5 min), Hellethionin D (18.3 ± 0.4 min), Hellethionin E1 and E2 (20.1 ± 0.6 min). Hellethionin D fractions were assayed for purity (see below), pooled, concentrated to 5 mg/mL and subjected once again to a preparative HPLC run with the same method. The final Hellethionin D fractions were collected at retention time of 17.8–18.8 min, assayed for purity (see below), pooled, and lyophilized.

Purity assay of the isolated hellethionins was performed on a Bio-Tek Kontron HPLC System 525 with DAD detector 545 and with a EC 250/4 Nucleosil 100-5 C8 column (Macherey Nagel, 200 mm length, 4 mm diameter) in a linear gradient from 5% buffer A to 85% buffer B in 40 min (buffer A: 0.1% *ortho*-phosphoric acid in water, buffer B: 100% acetonitrile).

Hellethionins were further identified by MALDI-TOF MS.

3.3.2 Crystallization

The crystallization of Hellethionin was carried out using a protein solution without further purification prepared from lyophilized protein (45 mg/mL in 20 mM HEPES pH 8.5). Hanging drop crystallization experiments were set up with Hampton VDXm pre-greased plates (0.6 mL reservoir), MD CrystalClene cover slips holding a drop of varying composition. The crystallization conditions had been derived from a hit in Hampton Crystal Screen (Condition 43: 40% PEG 3350, 0.2 M $LiSO_4$, 0.1 M TRIS pH 8.5) in several refinements. The different reservoir and drop compositions for the measured crystals are given below. The cryoprotectant solution contained a 1:1 mixture of reservoir and glycerol. The drop was mixed 1:1 with this solution, to yield 25% glycerol soaking for the crystals in the drop. Single crystals were mounted on MiTeGen MicroMounts and flash cooled by plunging into liquid nitrogen.

crystal	**xtal1**	**xtal2**	**xtal3**
reservoir	0.1 M TRIS pH 7.0,	0.1 M TRIS pH 7.0,	0.1 M BIS-TRIS pH 5.5,
(600 µL)	0.2 M $MgCl_2$, 1.9 M NaCl	0.2 M $MgCl_2$, 1.9 M NaCl	2.7 M NaCl
drop	1 µL reservoir	1 µL reservoir	1 µL reservoir
	0.8 µL protein solution	0.8 µL protein solution	1 µL protein solution
	0.2 µL 0.1 M NaI	0.2 µL 0.1 M glycine	
size	200 x 200 x 50 μm^3	150 x 180 x 45 μm^3	230 x 200 x 50 μm^3

3.3.3 Data collection and processing

Two data sets with high multiplicity were collected at DESY EMBL beam line X12 using a Marmosaic 225 CCD detector. One data set, xtal3, was collected at a Bruker Smart 6000 rotating anode diffractometer equipped with Incoatec multilayer optics and an Oxford cryo cooling system. As it was measured on a three-circle goniometer and the measurement took 18 days with several stops for de-icing the dehumidifier coil, three runs with $R_{int} > 25\%$ were removed from the data. In all cases, measurement temperature was maintained at 100 K. Data indexing and processing were accomplished with XDS (Kabsch, 2010), except for xtal3, which was processed with SAINT (Bruker, 2003). Scaling was applied with SADABS (Sheldrick, 2009). The space group symmetry of the tetragonal crystal was I422. The protein has 46 residues. Estimating a solvent content of 50% and an average amino acid residue volume of 140 $Å^3$, 8 to 9 protein monomers were assumed in the asymmetric unit.

3.3.4 Structure solution

The model with PDB code 1NBL was trimmed to residues 3 to 33 with side chains retained. The model is shown in Fig. 3.3 on page 39. Molecular replacement was attempted in a multi-solution PHASER (McCoy *et al.*, 2007) approach and successive SHELXE density modification (Sheldrick, 2010) on a grid of computers running CONDOR. This was achieved by using a modified version of ARCIMBOLDO (Rodríguez *et al.*, 2009). This version used the prepared search model instead of helical fragments generated *ab-initio*. After expansion, 36 putative sulfur atom positions were

determined using the merged and anistropically scaled data sets xtal1bc and xtal2ab. These were used for a new run of density modification and subsequent expansion in SHELXE.

3.3.5 SHELXE parameterization

Several parameters were tested for heavy atom search and tracing: Choice of data set, anisotropic scaling, given solvent content and for auto tracing NCS option, usage of anomalous scatterer positions and time factor. All searches for anomalous scatterers started with phases from the best trace by the modified ARCIMBOLDO-Version.

3.3.6 Refinement and validation

For calculation of R_{free}, 5% of the reflections were set aside. The experimental density generated by SHELXE (which was expanded to 1.0 Å by *free lunch* algorithm) was used for initial model building in COOT (Emsley *et al.*, 2010). All residues present in the final model could be built and mutated at this stage. The structure was refined with REFMAC (Murshudov *et al.*, 1997) against a data set merged from all data obtained from xtal1 (see Table 3.5 on page 44). During the final stages of the refinement, TLS rigid-body constraints (Winn *et al.*, 2001) were introduced. For this, each of the seven protein chains was defined as one domain. 318 residues are present in the final model. Several chloride and sodium ions were included along with the water molecules. The low average B factor of the solvent points to more water molecules being ion positions, which were not distinguishable. The general weighting scheme of geometric restraints against data in REFMAC was optimized testing different weights in a 100-cycles refinement cycles (to ensure convergence) by means of the best negative log likelihood gain (Tickle, 2007). The final model converged at an R factor of 19.0% (R_{free} = 22.1%). Quality checks of the final structure were performed using MOLPROBITY (Chen *et al.*, 2010). Refinement statistics are shown in Table 3.6 on page 44. No residues were observed in the generous and in the disallowed regions of the Ramachandran plot.

For calculation of r.m.s.d., the program LSQMAN (Kleywegt, 1996) integrated into a PYTHON script for automation was used.

3.3.7 Calculation of artificial data

The **PDB** of the final REFMAC refinement was converted into **INS** format with SHELXPRO (Sheldrick, 2008). The structure was then read into XPREP. Data sets with Friedel pairs, but uniform standard deviation, are generated by reading in a structure instead of data to XPREP automatically. The anomalous signals are added according to the wavelength defined by the user.

3.4 Results and discussion

3.4.1 Crystallization, measurement and data processing

The crystals were of uncommon ellipsoid-plate-like shape, typically less than 60 µm thin and 100-500 µm wide. Three data sets were collected at DESY X12 and copper-Kα home source. Diffraction data statistics are given in Table 3.1. The number in the data set name refers to the crystal, the letter to the data set (or synchrotron run) from this crystal. Scaling with SADABS (Sheldrick, 2009) was applied in all cases. The data set merged from all three synchrotron measurements which was used for MR and refinement is summarized in Table 3.5. The crystals had tetragonal symmetry (space group I422).

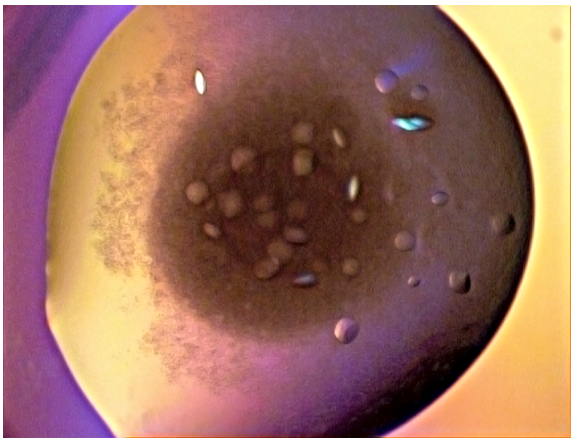

Figure 3.2: Ellipsoidal crystals from Hellethionin D

3.4.2 Structure solution

Despite a good anomalous signal, no suitable solution with SHELXD (Sheldrick, 2010) could be found.

Together with I. Usón, molecular replacement with the NMR structure as search model and PHASER was attempted. Several models were therefore generated from 1NBL by trimming side chains and the main chain by hand and with the program CHAINSAW (Stein, 2008). No successful MR solution could be gained.

Therefore, a modified version of ARCIMBOLDO (Rodríguez *et al.*, 2009) was used to do a multi-solution PHASER search using a CONDOR-run computer cluster and SHELXE (Sheldrick, 2010). The model which led to a successful solution (shown in Fig. 3.4) represented the two helices of the NMR structure connected by a loop region and 3 (of a total of 4) disulphide bridges (see Fig. 3.3). Only two models could be placed in the asymmetric unit, all other putative solutions were discarded in the PHASER translation search. This solution equals roughly 19.5% of all residues in the final model.

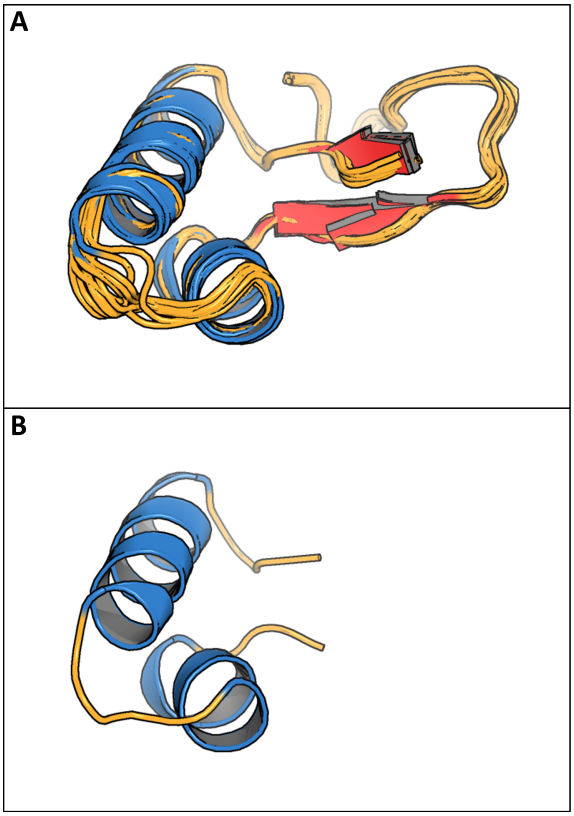

Figure 3.3: **A.** The NMR structure 1NBL (Milbradt *et al.*, 2003) **B.** The model which was used for successful solution.

Table 3.1: Summary of the data collection statistics. Values in parentheses refer to outer resolution shell.

	xtal1a	**xtal1bc**	**xtal2ab**	**xtal3**
source		DESY X12		SMART 6000
unit cell dimensions (Å)	a = 129.83, c = 103.99	a = 129.50, c = 103.56	a = 129.96, c = 105.02	a = 129.42, c = 102.25
space group		I 4 2 2		
wavelength (Å)	0.95400	1.90000	1.90000	1.54178
oscillation range	180°	360°	360°	4-circle diffractometer
resolution range (Å)	24.97–1.742 (1.84–1.74)	25.01–2.018 (2.13–2.03)	24.13–1.98 (2.08–1.98)	34.10 –2.70 (2.80–2.70)
no. of observations	600695 (67592)	1036510 (87398)	298321 (72762)	1241876 (100353)
unique*	45243 (6602)	29050 (4039)	26170 (3362)	12269 (1008)
multiplicity*	13.20 (9.92)	35.54 (21.65)	26.65 (21.64)	99.73 (85.12)
completeness* (%)	99.4 (96.9)	99.6 (97.8)	99.9 (100.0)	98.5 (85.5)
Friedel-completeness (%)	99.0 (94.3)	99.5 (97.0)	100.0 (100.0)	98.6 (85.12)
R_{int} (%)**	8.65 (60.65)	14.72 (81.44)	16.87 (75.69)	15.21 (35.75)
R_{rim} (%)**	9.00 (63.88)	14.93 (83.15)	17.20 (77.45)	15.28 (35.93)
R_{pim} (%)**	2.43 (19.50)	2.47 (16.53)	3.32 (16.32)	1.51 (3.57)
Mean I/σ(I)	21.32 (2.24)	29.89 (4.00)	21.58 (3.06)	38.45 (17.39)
d"/σ(d")	0.76 (0.72)	1.20 (0.73)	1.09 (0.83)	1.08 (0.86)
R_{anom}**	5.75 (68.8)	8.76 (68.25)	11.25 (64.32)	4.33 (9.05)

*For Friedel pairs merged.

**As defined in the appendix on page on page 85.

The correct solution could be discriminated by the automatic chain expansion in SHELXE (CC against native data of 37.8%, average chain length 39.1). Rerunning SHELXE for more cycles or running only 5 cycles but correcting the solvent content from 0.45 to 0.55 (7 rather than 8 molecules) improved the CC against native data to 43.8% and the average chain length to 44.3. (This step was optimized as given in section 3.4.4.)

Putative sulfur atom positions were determined from this trace and the rest of the structure was discarded. By this means, the MR solution had "bootstrapped" the SAD phasing. Only the derived 49 sulfur positions were used for a new run of density modification and subsequent expansion in SHELXE. The final trace contained 299 of 322 residues, with 16 misplaced terminal residues present. The structure was subjected to the PISA Webservice (Krissinel & Henrick, 2004, 2007). No symmetry relation could be found between the seven molecules in the ASU. The biggest inter-molecule surface is 464 Å^2 of 3100 Å^2 total molecular surface, which indicates the molecules are biological monomers.

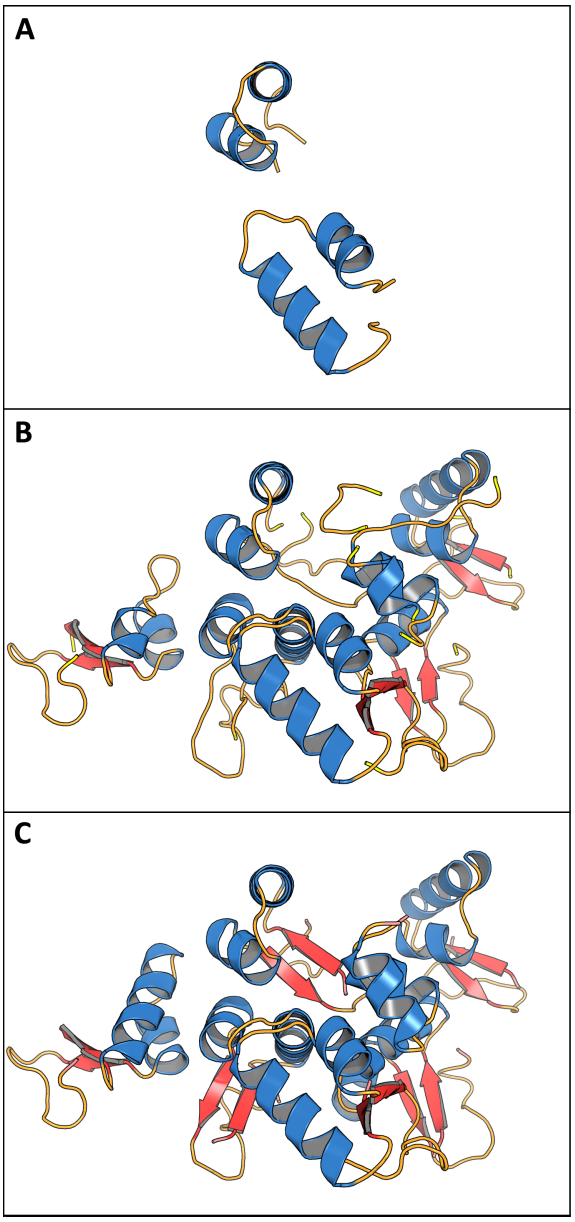

Figure 3.4: **A.** Two models placed by PHASER. **B.** ARCIMBOLDO trace. **C.** Trace after SAD.

3.4.3 Initial failure of molecular replacement

Only by successful structure solution it becomes evident whether a model was good enough for MR solution or a measured anomalous signal was sufficient for S-SAD.

Regular molecular replacement failed for this structure. The r.m.s. (C_α) deviation of the model to the final crystal structure was 1.34 Å. This is already in the "twilight zone" for MR models according to Chothia & Lesk (1986). Later, it could be shown that PHASER could solve the phase problem easily with the X-ray structure of viscotoxin A1 (r.m.s. C_α deviation 0.76 Å). Ironically, this structure had been determined in our lab by means of S-SAD.

With the NMR structure as search model, PHASER gave many potential solutions, among them the one with only 19% of all amino acid residues placed in the ASU (two copies) which led to structure solution. Testing as many solutions as in this case requires much computational power – here the CONDOR grid and a robust processing framework like the one of ARCIMBOLDO.

Nonetheless, by means of MR-SAD, we could utilize the NMR structure to gain a structure solution. With this, we provide another example of phasing employing an NMR structure (Chen *et al.*, 2000). SHELXE played a key role in this as the indicator of a good MR solution and it provided additional phase information through density modification. The high solvent content (61%) favoured density modification and tracing to a high completeness. Although phasing was not feasible by the established methods we employed and the NMR struc, this multi-solution MR-SAD method resulted in an almost complete, model-bias free trace of all seven protein chains.

3.4.4 SHELXE parameterization

SHELXE (Sheldrick, 2010) was tested for the best parameters to find the heavy atom positions and the best tracing method. All tests started with the phases obtained from the modified ARCIMBOLDO run, i.e. the trace of the MR solution. CC refers to the CC against native data; AA to the average chain length.

The data set with the highest anomalous signal as well as a merged data set of xtal1bc and xtal2ab was employed with and without anisotropic scaling (a.s.).

Table 3.2: Tests for data set choice with and without anisotropic scaling. For the sulfur search, the command **`shelxe XX.pda YY -m50`** was used, for tracing **`shelxe XX YY -m50 -a5 -q -s0.45 -e1 -l3`** plus options given in the table.

data set	**xtal1bc**		**xtal1bc, a.s.**		**xtal1bc+2ab**		**xtal1bc+2ab, a.s.**	
revised atoms (found)	32 (46)		35 (45)		49 (49)		39 (45)	
	CC	**AA**	**CC**	**AA**	**CC**	**AA**	**CC**	**AA**
no additional options	41.75%	34.3	42.61%	31.0	41.65%	45.0	42.78%	37.9
-h[no. revised]	38.04%	28.3	37.96%	31.3	37.93%	39.0	37.43%	34.1
-h[no. revised] -n7	41.75%	34.3	42.61%	31.0	41.65%	45.0	**43.81%**	**44.3**

The ASU contained 56 sulfur atoms in total; the revised atom positions were not checked for false positives after structure solution ("revised" here refers to the SHELXE output). Data sets scaled anisotropically with XPREP (Sheldrick, 2011) clearly gave better tracing in comparison. Also, it could also be shown that the merged data of xtal1bc and xtal2ab yields more marker atom positions and a better trace than the data set with the highest anomalous signal (xtal1bc) alone. This is because errors are reduced by merging data of comparable quality from different measurements and different crystals.

It could also be shown that including the revised heavy atom positions (**`-h`**) and the new NCS option (**`-n`**), which uses the similarity between several copies of the same protein in the ASU, improves the auto tracing. Only for the merged and anisotropically scaled data, NCS could be found by SHELXE.

To test the influence of the solvent content, the anistropically scaled data from xtal1bc and xtal2ab were used.

Table 3.3: Tests for the best solvent content in SHELXE. For auto tracing, the substructure from the 70% solvent content substructure search was used (substructure search command: **`shelxe XX.pda YY -s[fraction] -m50`**, tracing command: **`shelxe XX YY -m50 -a5 -q -s[fraction] -e1 -l3`**).

solvent content	revised (found)	CC	AA
45%	41 (43)	39.95%	33.6
50%	43 (46)	40.06%	38.6
55%	43 (45)	40.50%	38.4
60%	43 (45)	40.92%	43.3
65%	45 (45)	39.70%	34.6
70%	47 (48)	39.16%	33.2
75%	44 (46)	40.03%	33.4

The solvent content was adjusted (**`-s`**). The real content of disordered solvent in the crystal was 61.4%. For the anomalous scatterer search, a slightly overestimated (70%) value yielded the best result, while for tracing the value next to the real solvent content was best (60%).

The influence of the time factor for helix and peptide searches was tested.

Table 3.4: Tests for SHELXE auto tracing time factor. The anistropically scaled data from xtal1bc and xtal2ab were used (command: **`shelxe XX YY -m50 -a5 -q -s0.65 -e1 -l3 -h47 -t[N]`**).

time factor	1 (default)		2		3		4		5	
	CC	AA	CC	AA	CC	AA	CC	AA	CC	AA
	39.70%	34.6	39.26%	28.0	40.02%	27.7	39.41%	36.9	39.55%	33.6
time factor	**6**		**7**		**8**		**9**		**10**	
	CC	AA	CC	AA	CC	AA	CC	AA	CC	AA
	38.95%	33.2	39.76%	28.3	39.83%	37.1	39.99%	37.9	39.57%	34.2

Changing the time factor for the tracing ("-t") did not significantly improve the CC against native data or the average chain length, although it proved useful in other cases (Sheldrick, personal communication).

3.4.5 Refinement

A SHELXE *free lunch* map extended to 1.0 Å resolution was used for the initial model building from the backbone. The trace of all seven chains in the ASU was refined further with REFMAC. As the B factors for solvent waters were very small, and some of them showed peaks in the anomalous density map, 35 of 'waters' were assigned to Cl^- and 28 to Na^+ (in some cases with halved occupancies), depending on charge and coordination surroundings of the individual positions. This was justified as the crystals had grown out of high salt content conditions (compare crystallization conditions in Table 3.3.2). The low average B factor for solvent waters still hints to more ions among them. For the final structure, no residues were observed in the generously allowed and in the disallowed regions of the Ramachandran plot.

Table 3.5: Summary of data statistics for the merged data; values in parentheses refer to outer resolution shell.

data statistics for merged data set	
unit cell dimensions (Å)	a = 129.829, c = 103.994
space group	I 4 2 2
wavelength (Å)	0.954/1.900/1.900
resolution range	1.95-25.01 (1.95-2.05)
no. of unique observations	32570 (4446)
multiplicity	45.90 (16.75)
completeness (%)	99.9 (99.9)
R_{pim} (%)**	1.73 (9.22)
mean I/ σ(I)	38.37 (6.38)

** As defined on page on page 85.

Table 3.6: Refinement statistics.

Refinement statistics	
resolution range (Å)	25.01–1.95
reflections (working set)	30205
completeness (working set, %)	99.9
reflections (test set)	1622
solvent content (%)	61.39
no. of protein atoms	2319
no. of water molecules	381
no. of ions	28.5
protein molecules per ASU	7
R (%)	18.9
R_{free} (%)	21.9
average B factors (Å^2)	
overall	27.7
protein atoms	28.44
waters and ions	23.45
r.m.s.d. from ideal geometry	
bond lengths (Å)	0.010
bond angles (°)	1.132
Ramachandran plot, residues*	
in most favoured regions (%)	98.68
in allowed regions (%)	1.32

*Calculated with MOLPROBITY

3.4.6 Comparison with the NMR structure

Comparison between the NMR ensemble (20 chains) and the seven chains of the X-ray model gave an r.m.s.d. of 1.9 Å (for all protein atoms) and a generally similar fold. Other than in the NMR structure, the multi-threonine loop 36–39 was not folded differently than in the other thionins, as shown in the next section.

3.4.7 Comparison with other structures

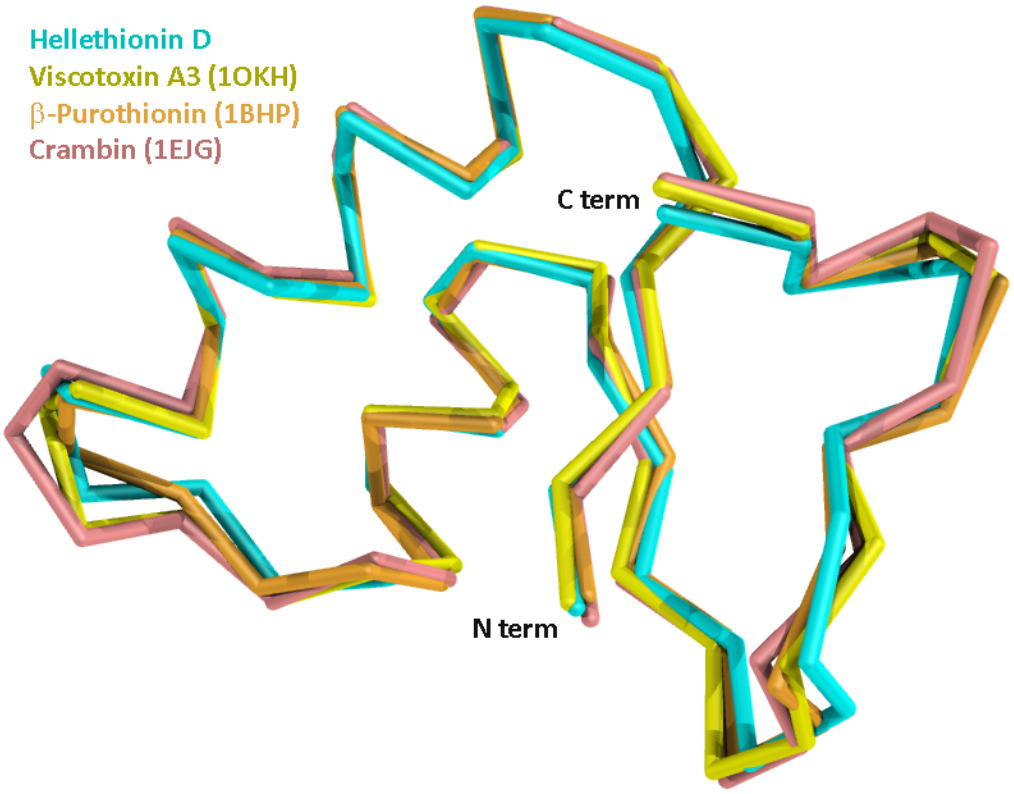

Figure 3.6: Main-chain overlay between chain E of the Hellethionin D X-ray structure (representing the common fold of the protein molecules in the ASU, cyan) and the related structures mentioned below.

Three-dimensional alignment with DALI (Holm & Sander, 1997) shows that the fold resembles those of other thionins, with Viscotoxin A3 yielding the highest Z-score:

Name	PDB	Z-score (DALI)	r.m.s.d. (DALI)	Sequence identity
viscotoxin A3	1OKH	8.1	0.7	54%
β-purothionin	1BHP	7.9	0.8	46%
crambin	1EJG	7.8	0.9	30%

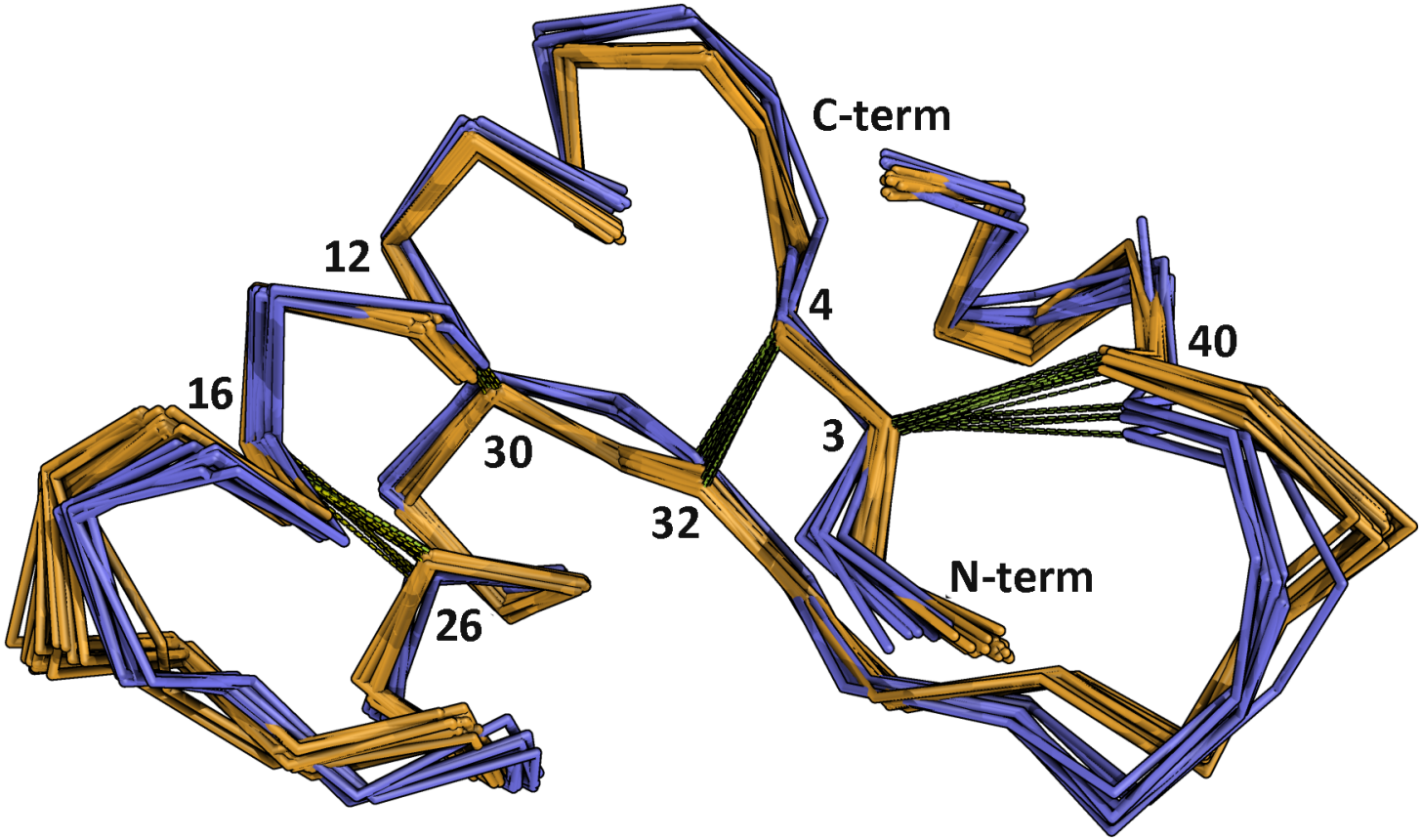

Figure 3.5: Comparison between the 7 conformations found in the X-ray structure (blue) and the NMR ensemble (20 conformations, ocre). Cysteine residues are numbered. The multi-threonine loop 36–39 is in the lower right corner and clearly differs between the models.

3.4.8 NCS and crystal structure pores

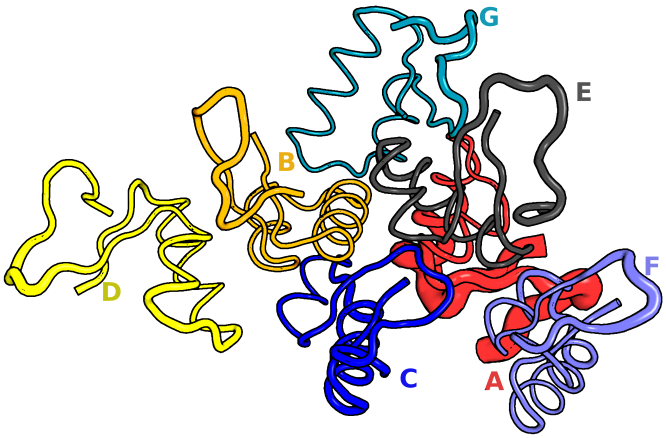

Figure 3.7: The seven chains in the ASU as B factor putty. Note the high B factors in chain A.

Chain A was poorly resolved in the electron density and its disulphide bridges had a weak signal in the anomalous electron density map. The molecule was very flexible with high B factors (compare the B factor putty given in Fig. 3.7). Modelling as disorder, with analogous fragments from the NMR model or from the other six copies in the asymmetric unit did not improve the density fit. Finally, seven residues of this chain were missing in the density. This was found to be due to solvent exposure: Eight copies of chain A, related by crystallographic symmetry, form a pore in the crystal structure. The diameter of the pore is roughly 35 Å across with a special position (Wykoff letter **a**, site symmetry 422) is lying in the middle, as which is depicted in Fig. 3.8.

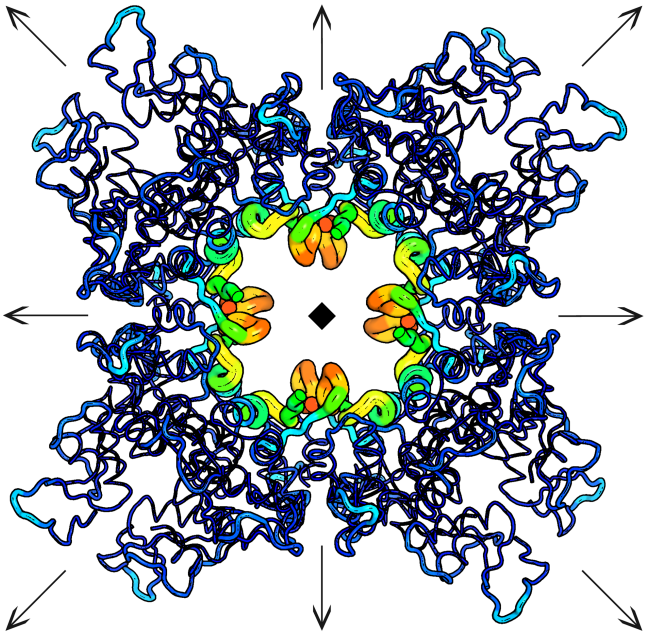

Figure 3.8: View along the fourfold axis of crystal packing pore, which is surrounded by the A chains from 8 asymmetric units. The protein chains are shown as B factor putty.

3.4.9 Data analysis

With the structure of Hellethionin D solved by MR-SAD, it remained unclear why it could not be solved by S-SAD alone in the first place. A thorough data analysis was carried out. We analyzed therefore the measured data sets (see Table 3.1) using these indicators:

- correlation of the data sets
- correlation of the anomalous signal between the data sets
- R_{anom}
- $d''/\sigma(d'')$

Plots of these indicators against resolution similar to the ones produced for some indicators by HKL2MAP (Pape & Schneider, 2004) were desirable. XPREP was modified to plot data quality indicators graphically. We also generated ideal data were calculated from the final structure of Hellethionin. These contained the anomalous signal ($\lambda = 1.9$ Å).

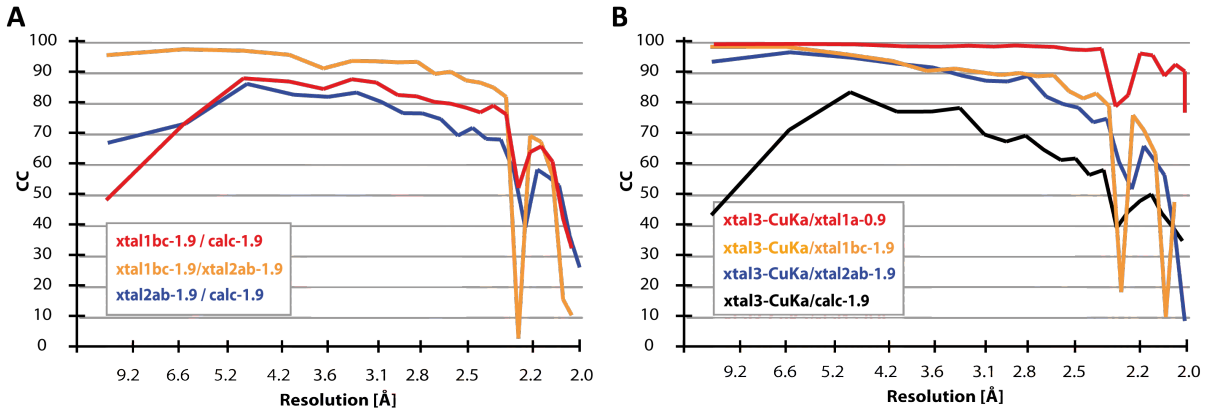

Figure 3.9: Correlation coefficients against resolution. **A.** Artifical data and synchrotron data sets. **B.** CC against xtal3, measured with a Cu-Kα home source.

The correlation coefficients of the data sets plotted in Fig. 3.9 show interesting features. Data sets xtal1bc and 2ab (A, yellow curve) are in good agreement, which explains why the merged and anistropically scaled data of these gave the best results in the SHELXE parameterization tests. The correlation with the artifical data (λ=1.9 Å) is not very good in the low-resolution region. For the data set xtal3, given in Fig. 3.9 B, which was measured at our home source, the correlation with the synchrotron data is good. We could not find a suitable explanation for the sharp drop in correlation at 2.3 Å. Again, the correlation with the artificial data is low.

From the plot of $d''/\sigma(d'')$ it becomes clear that the anomalous signal was sufficiently strong. As for the artifical data, the uncertainty is not given, $d''/\sigma(d'')$ was not calculated. R_{anom} is surprisingly low above 6.9 Å resolution for all data sets, while d“ is high, as expected, except for the artificial data.

The anomalous correlation between data sets is shown in Fig. 3.11. We also evaluated the anomalous self-correlation, which is not shown. It was generally very good. The synchrotron-measured data sets correlate well with each other in their anomalous signal. The correlation of them with the in-house data is slightly worse, what is to be expected, since the anomalous signal of sulfur is weaker at 1.541 Å wavelength. The correlation with the artifical data is worse, especially in the low resolution range.

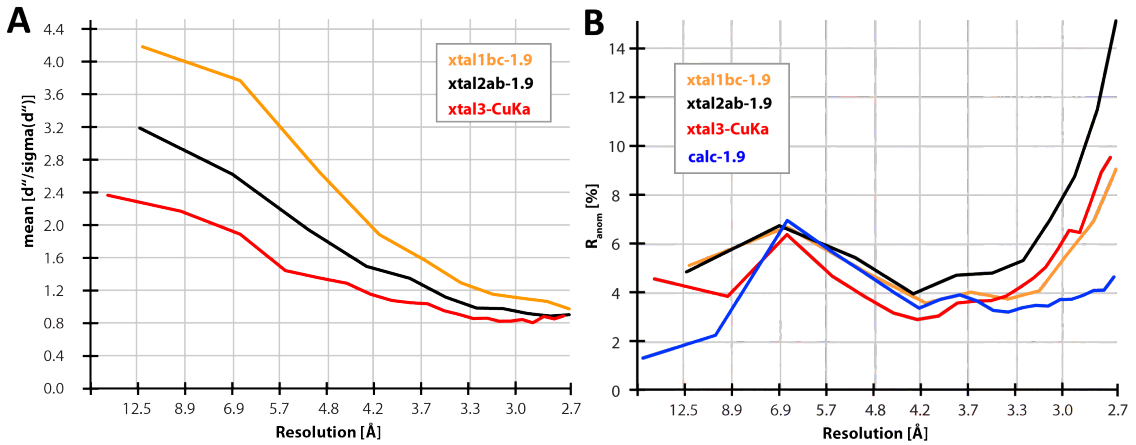

Figure 3.10: Anomalous data indicators. **A.** d″/σ against resolution shows a strong anomalous signal. **B.** R_{anom} against resolution.

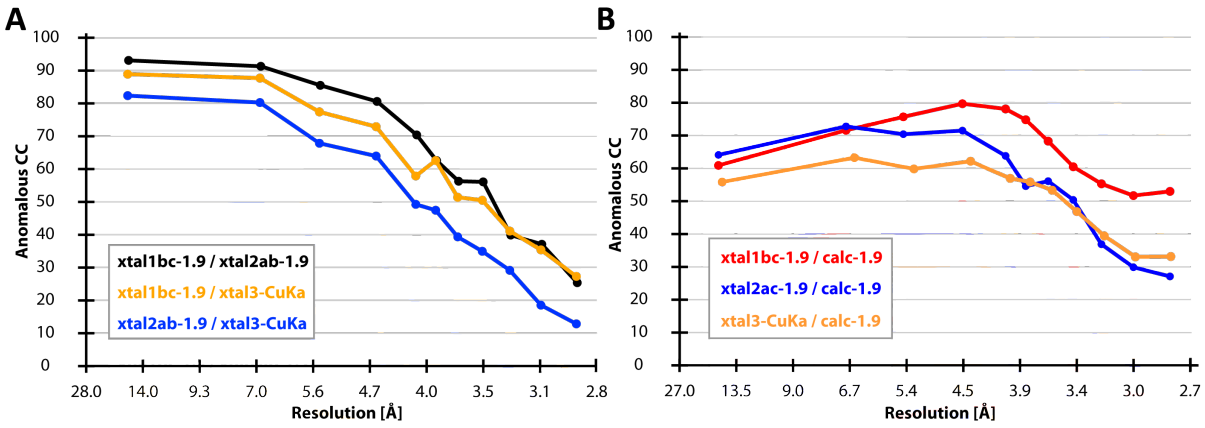

Figure 3.11: Anomalous correlation coefficient **A.** Between measured data **B.** Between measured and artifical data calculated for $\lambda = 1.9$ Å.

From these data statistics we could not properly determine why S-SAD was not possible directly. We considered that the bad correlation and high R_{anom} for the artifical data in the range lower than 6.5 Å resolution was linked to this.

3.4.10 Poor correlation of artifical data

One possibility for the bad anomalous signal and the poor correlation of the artifical data would be a low completeness in the inner shells. If not all reflections are measured, this might influence R_{anom} strongly. The completeness in inner shells (with only symmetry mates merged) was checked and found to be 99.5% – 100% up to 20 Å resolution. Because of the smaller beam stop at our in-house source, the data completeness in the resolution shell 20 Å – 30 Å was 82.6% for xtal3. The anomalous signal decreases from 6.5 Å on, hence the completeness seems not to be the reason for the discrepancy.

The effect could also be caused by disordered halide ions in the solvent pores of the crystal (see Fig. 3.8), as Hellethionin D had been crystallized from solutions with a high salt content

(compare Table 3.3.2). The high B factors of these ions would result in low resolution anomalous scattering only, which might obscure the anomalous signal from ordered anomalous scatterers. However, the effect should be significantly lower in the data sets xtal2ab and xtal3, for there was no iodine present in the crystallization mixture. This is not the case.

The most likely explanation for the poor correlation with the artificial data led us to one of the central problems in protein crystallography: The disordered solvent. This "soup" of water, ions and other compounds is not distributed completely random in the cell. Therefore, it scatters X-rays in such a way that interference occurs. Both the phase as well as the intensity of the reflections are affected. As XPREP does not use any solvent model to make up for the disordered solvent regions, errors in the artificial data are generated. This was the reason why the artificial data sets do not adequately model the low resolution anomalous signal. Due to their high B factors, disordered solvent regions are only having a significant influence on the low resolution phases and intensities. As long as we have no proper solvent model, we will not be able to explain low resolution anomalous scattering.

3.5 Outlook

Since software and compatibility improve, we can now freely combine methods and phase information to push boundaries for what can be phased in protein crystallography. We combined weak phase information from different sources and phase improvement in SHELXE to give a solution, where S-SAD and conventional MR with the NMR structure alone failed and obtained the structure of Hellethionin D without any model bias.

However, we cannot clearly explain why the substructure can not be found by SHELXD in a conventional S-SAD approach. Comparison with artificial data suggests that better knowledge of the disordered solvent regions might lead us to an answer – the solvent strongly influences the low-resolution reflections. This is not a limitation, but a chance: Exact experimental phases for these reflections might be gained from highly accurate MAD structures, and their intensity is measured in our experiments. As we know the differences in both phase angle and intensity between them and the ones to be expected from our model, we might be able to determine a new solvent model from this.

However, for now, we still seem not to understand the nature of the anomalous signal in combination with the solvent well enough. While the high solvent content helps for density modification, it might also be the reason why we could not solve the data with SHELXD alone initially. After successful phasing with MR-SAD, we used the correct number of anomalous scatterers in SHELXD, but up until now experimental phasing of the data without bootstrapping by MR has not been possible.

The evaluation of the phenomenon led to the development of the tool ANODE, which will be discussed in detail in the next chapter.

4 ANODE: Validation with anomalous density

4.1 Introduction

The program ANODE („ANOmalous DEnsity") was initially developed to clarify the role of anomalous scatterers in Hellethionin D. It uses experimental data to give anomalous density peaks and the averaged anomalous signals per atom type for a given input model.

The program proved to be very useful, not only for MR-SAD, but also for validation in experimental phasing as well as to assess data and models. In this chapter the program's functionality and parameterization are discussed.

4.2 Program description

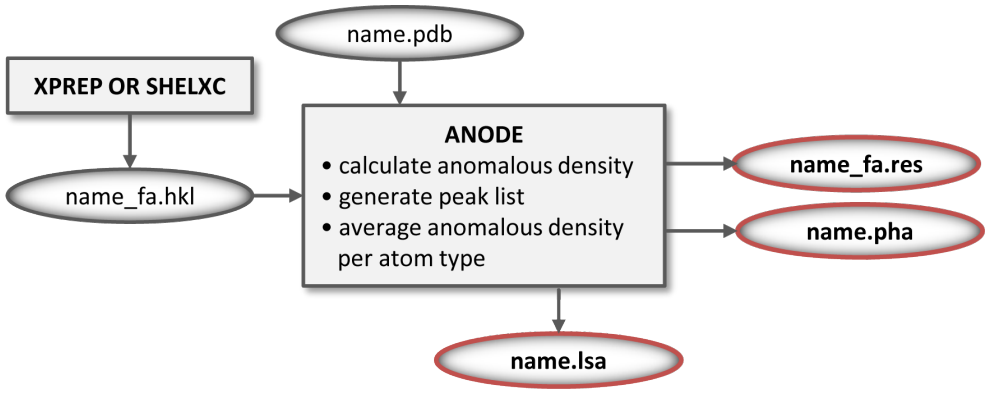

Figure 4.1: Data flow for ANODE.

ANODE reads an **`name_fa.hkl`** file from XPREP (Sheldrick, 2011) or SHELXC (Sheldrick, 2010). This file contains $|F_A|$ and its uncertainty $\sigma(|F_A|)$, the marker atom contribution to the structure factor as well as as the phase shift α. It also reads a **`PDB`** file with a model, which does not need to contain marker atoms. Structure phases (without Bijvoet differences) ϕ_P are calculated from the model. α from the **`name_fa.hkl`** file is then subtracted from these to get substructure phases ϕ_A. To calculate the so-called anomalous map (in the case of SAD or MAD), the required amplitudes $|F_A|$ are also obtained from this file. The result is a table with averaged anomalous density values or, optionally, for each and every atom. The map is in **`PHS`** format and can be displayed *e.g.* in COOT (Emsley *et al.*, 2010). A unique peak list is generated, where the interpolated highest peaks of anomalous density and the nearest neighbour atoms are given. A **`name_fa.res`** file with those is generated for usage in SHELXE. As element type for these positions either the heaviest atom type is chosen, or, if that would be chlorine in the presences of sulfur, sulfur.

4.3 Parameterization

4.3.1 Available options

ANODE has a number of command line options:

If the **`name_fa.hkl`** could be indexed differently within the space group given by the **`PDB`** file, the program gives a warning and the alternative indexing option (**`-i`**) can be used. For the space groups P3, $P3_1$ and $P3_2$ four indexing possibilities exit, which can be chosen by a number.

The anomalous signal does not extend to the scattering limit. A maximum resolution for F_A can be regulated by a sharp cut-off (**`-d`**) or by damping (**`-b`**).

The program prints anomalous densities averaged by atom name and residue type. But the number of atom types can be limited (**`-m`**) or the anomalous density can be given for every atom in the PDB without averaging (**`-a`**). The peak list (in **`name_fa.res`**) can be regulated by the minimum height relatively to the strongest peak (**`-t`**) and by the maximum number of peak output (**`-h`**), whichever is lower. Finally, the resulting map's accuracy can be regulated (**`-r`**) by adjusting the factor for maximum h, k and l for the Fast Fourier Transform grid, for example for figure creation. The program version discussed here uses the default options **`-b4.0 -d1.0 -h80 -r5.0 -t0.15.`**

In general, experimental phasing aims for a high contrast between marker atom substructure and noise. Consequently, the peak height is a general indicator of a good choice of options.

4.3.2 Resolution vs. B factor

It is common practice to cut the outer resolution shells in substructure search and refinement, because the signal-to-noise ratio – as it is only a fraction of the whole measured value – is often too low. It is also argued that at lower resolution, disulphide bridges and disordered marker atoms fuse into single peaks.

ANODE allows for cutting the resolution. But also, a B factor can be applied to the outer resolution range, dampening the high-resolution data of which the accuracy often suffers from low signal-to-noise ratios. This is also a feature of SHELXE (Sheldrick, 2002), where the B factor is not tuneable and set to 4.0. An interesting question is whether the substructure is improved by different B factor settings and resolution cut-offs, or a combination thereof. The test results are shown in Fig. 4.2 on the facing page.

One result is that only cutting the resolution does not improve the average peak heights for the anomalous scatterers significantly. B factor tuning is more effective to heighten the anomalous signal than a crude resolution cut-off. The rather high B factors between 16 and 25 showed the best result for Hellethionin D. The same test was applied to human RNase T2 (data not shown). Here as well, a resolution cut-off did not significantly improve the anomalous peak height – while B factors over 15 did.

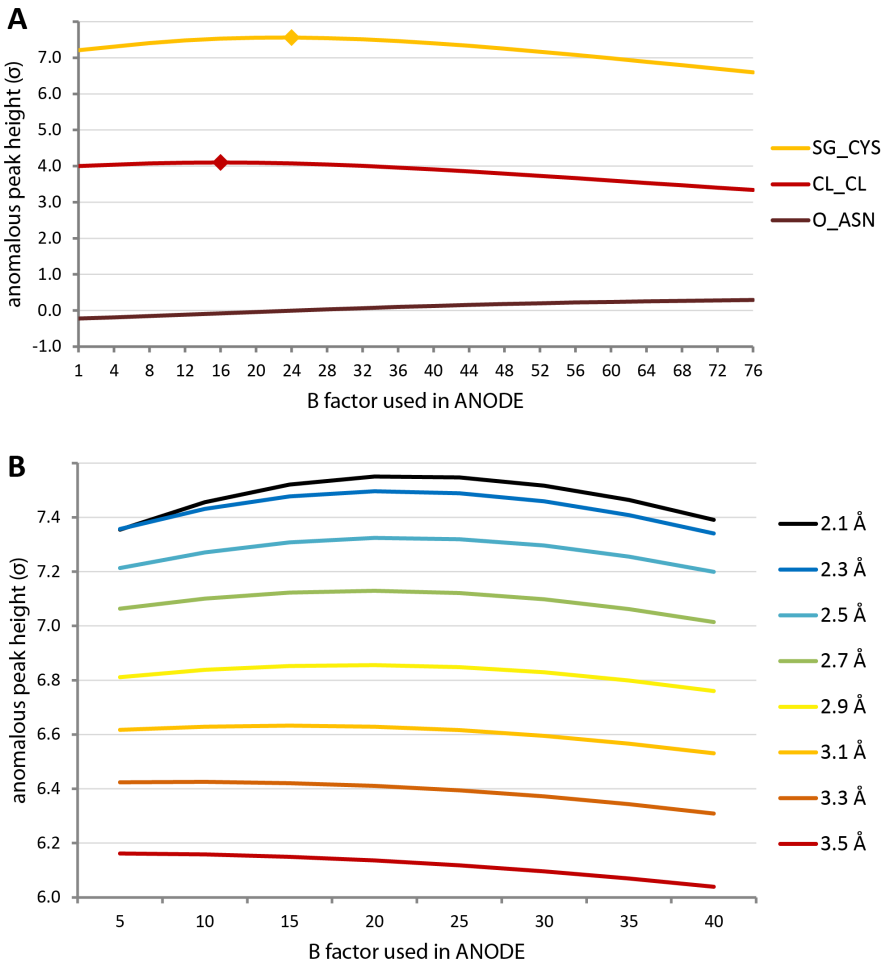

Figure 4.2: Input B factor against output peak height in ANODE. **A.** Three peaks from Hellethionin D at native resolution. Cysteine sulfur (SG_CYS) and chloride (CL_CL) become higher with increased B factor. The maximum is marked. Note that the peak of asparagine oxygen, which should not have a significant anomalous signal in this data set, increases with higher B factor. **B.** Combination of resolution cut-off and B factor in Hellethionin. The highest signal is achieved here with the best resolution (2.1 Å) and B at 20–25.

Dampening the outer resolution shells means not to remove the high resolution anomalous signal, but to weight it down in the calculation of the anomalous electron density. This test might be a hint that in experimental phasing, an absolute resolution cut-off is not optimal. The dampening of high-resolution data should be tested with a higher B factor, if not be tuneable in general.

4.4 Applications

4.4.1 Data set choice

Many quality indicators exist for data and the anomalous signal in particular (see section 5.6 on page 85). The peak height in the anomalous map is one of the most immediate of them. After all, these peaks define the marker atom substructure. Therefore, ANODE can be used to judge the anomalous signal in a given data set, with a suitable model at hand.

First, we probed the overall functionality and default parameterization of ANODE on the Hellethionin D data sets (data set statistics are given in Table 3.1 on page 40) and the final structure. All options were set to default values. As initial quality indicator, the average peak heights of cysteine sulfur and solvent chlorine atoms, given in standard uncertainties σ of the electron density, were used:

Table 4.1: Different Hellethionin D data sets used to calculate average peak heights in ANODE.

Hellethionin D						
command	**SG_CYS**	**CL_CL**	**O_TYR**	**d"/σ(d")**	**R_{anom}**	λ
`anode xtal1bc`	7.306	4.036	-0.549	1.20	8.76%	1.90000
`anode xtal2ab`	5.911	2.132	0.371	1.09	11.25%	1.90000
`anode xtal3`	2.066	0.868	0.493	1.08	4.33%	1.54178

From d"/σ(d"), the average peak height (in σ) of cysteine sulfur and of chloride it becomes evident that the data set xtal1bc had the strongest anomalous signal. The averaged anomalous electron density of tyrosine oxygen is given for comparison to signify noise. While xtal1bc and xtal2ab show a relatively high signal compared to tyrosine oxygen, xtal3 only shows an anomalous signal four times as high.

We used different human RNase T2 data sets (overall data statistics in Table 2.3 on page 22) for a similar test, also with the final structure model. For RNase T2, the data set rnase3_ds3 gives the highest anomalous peak. The data set rnase32 was the merged from rnase2_ds2 and rnase3_ds3 and in this case, seems not optimal to find the anomalous substructure, as its peak heights are relatively low.

Table 4.2: Different human RNase T2 data sets used to calculate average peak heights in ANODE.

Human RNase T2				
command	**SG_CYS**	**CL_CL**	**d"/σ(d")**	λ
`anode rnase2_ds2`	4.647	3.270	0.88	1.95000
`anode rnase3_ds3`	7.701	5.984	0.89	1.95000
`anode rnase32`	6.496	3.891	1.15	1.95000

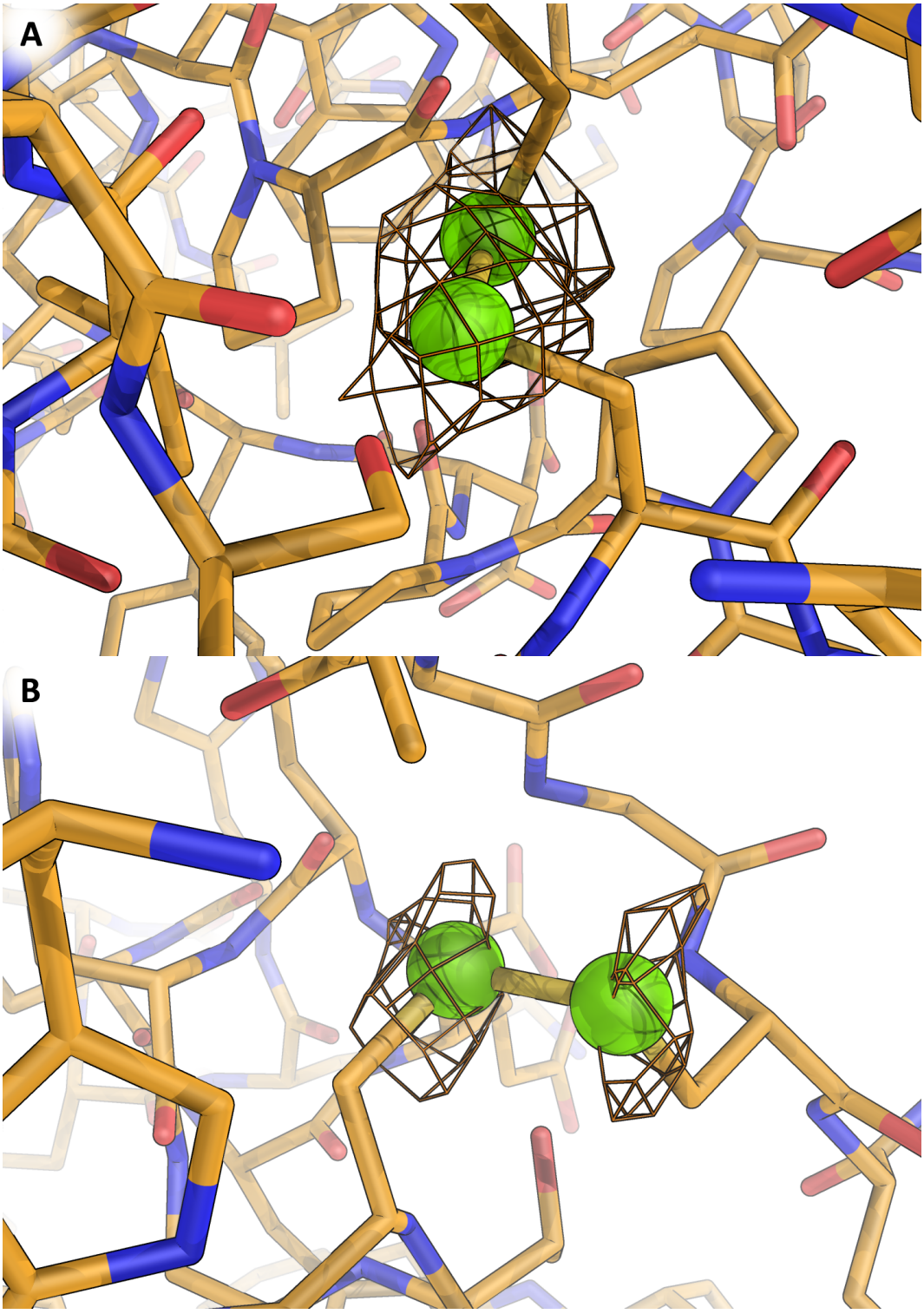

Figure 4.3: Anomalous density around disulphide bridges for human RNase T2. **A.** The disulphide bridge (48/55) is well defined in the density, as its position within the protein possibly protects it from radiation damage. **B.** This disulphide bridge (75/121) has suffered from radiation damage, resulting in breakage of the bond and potential loss of sulfur, resulting in a lower occupancy. Map shown at 2.2σ.

4.4.2 Validation

With ANODE, the position of marker atoms in a structure can be confirmed. Especially in isomorphous replacement methods, where the heavy atom map often results from another crystal with different unit cell dimensions, this can be very useful. As the cell is read in from the **PDB**, the peak positions will be automatically scaled to the model in the **PDB** file, making validation much easier. It can also serve as an easy way to get anomalous maps for figures, as shown in Fig. 4.3 and Fig. 4.4.

In the anomalous density map for human RNase T2, radiation damage became visible, as shown in Fig. 4.3. While unaffected disulphide bridges show a high peak, enclosing the disulphide ("super-sulfur"), disulphide bridges cleaved by radiation damage show separated, weaker peaks. By this means, ANODE can help analysing radiation damage in anomalous scatterers as well.

4.4.3 Input model choice and MR-SAD for Hellethionin D

ANODE can be used for MR-SAD, as it can read in an MR model and anomalous data prepared with SHELXC and write out the anomalous substructure, which can then be used in SHELXE. It then takes the role of SHELXD in conventional experimental phasing.

ANODE was tested with different input models to find the anomalous substructure. The calculated positions were compared manually to the 91 sulfur and chloride positions in the final structure using COOT (Emsley *et al.*, 2010). For all tests, the data set xtal1ab was used. The name_fa.res file was subjected to SHELXE density modification and auto tracing (command: **`shelxe XX YY -m50 -a5 -q -s0.45 -e1 -l3`**).

Table 4.3: MR-SAD with ANODE (command: **`anode -b20 name`**). The marker atom positions in the name_fa.res file were by default 80.

input PDB	highest peak (σ)	correct output positions	CC	AA
MR solution	4.713	12	6.66%	7.92
ARCIMBOLDO trace	9.905	54	31.93%	44.0
optimized trace	8.283	51	31.70%	33.9
final structure	12.273	60	32.10%	28.5

The optimized trace yielded lower peak height and fewer correct positions than the one given out by ARCIMBOLDO, giving evidence for more phase error resulting from this model, as well as the lower correlation coefficient against native data (CC) and average chain length (AA) in SHELXE. It also becomes clear that the MR solution alone would not have been accurate enough for MR-SAD – only 12 correct positions are not enough for a successful trace in SHELXE, which is indicated by a low CC against native data. There is good correlation between the maximum peak height and the number of correct marker atom positions, and hence, the quality of the anomalous substructure in name_fa.res.

4.5 Discussion and outlook

It was demonstrated that ANODE is a useful tool: It can be used to confirm marker atom positions and visualize radiation damage. The program also gives a good indication of the mean phase error of the model employed and the quality of long-wavelength data sets. The functionality of ANODE is not new, but it is easy to use, with a clear data flow. In this chapter we discussed mainly its application to SAD it can also be applied in a similar fashion for other experimental phasing methods. It can be used for MR-SAD, where it replaces SHELXD in elucidating the substructure. It might be of help in the development of experimental phasing and data processing methods.

The program supplements experimental phasing in SHELX well and can be easily automatized. ANODE has been distributed as beta test version to the SHELX community.

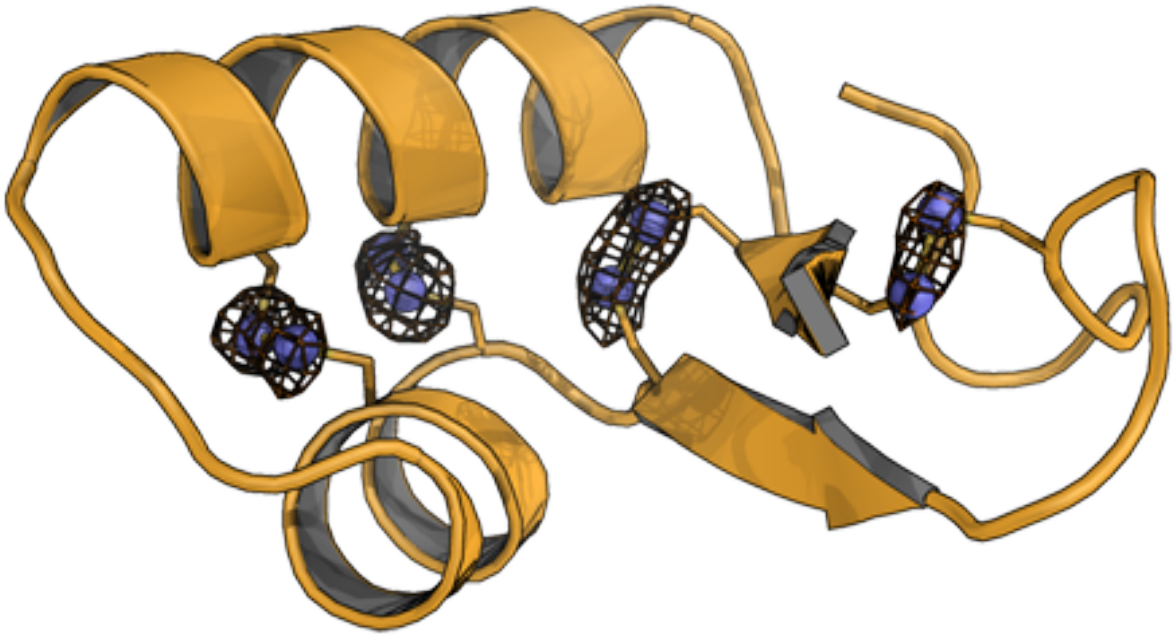

Figure 4.4: Anomalous density for one molecule of Hellethionin D. Generated with ANODE using data set xtal1bc and the final structure.

5 REST: Rigid-bond restraints in SHELXL

5.1 Introduction

In the routine solution of small molecule structures, an R-value over 10% is considered not acceptable for publication. The R-value is, roughly put, a measure of how good the model fits the data. Protein structures hardly ever have R-values under 10%, even if taken from well-ordered crystals and refined with optimal methods (Rupp, 2009). This leads to the conclusion that our modelling of protein crystal structures is lacking compared to that of small molecules. Several reasons for this discrepancy have been given: Proteins are big, flexible and pack with much more space between each other than small molecules. Hence, big regions in the crystal consist of disordered solvent which cannot described very well. Also, atomic displacement from an average position, vibrational and disorder behaviour is not yet fully understood for macromolecules. The TLS description of displacement is widely used and lowers the R_{free} significantly. Nonetheless, it is not clear why it works so well, as the assumptions on which TLS was developed do not hold strictly for proteins. For example, they might not move as rigid bodies. Also, there are certain drawbacks to TLS, which are discussed in detail on the following pages. In this project, we aimed to develop a method for the refinement program SHELXL (Sheldrick, 2008) which could lead to better modelling of the displacement of macromolecular structures, and compare it to the methods already available.

5.2 Background

5.2.1 Refinement

Structure refinement means fitting a model to observed data (the reflections) by adjusting parameters. For this fitting, the model can be regarded as a multi-parameter function which defines the relation between the model and the data. The two commonly employed target functions are maximum likelihood and least-squares. The latter can be seen as a special case of the first, where all data are distributed as Gaussians.

The data is weighted by its uncertainty σ and then, parameters are adjusted to minimize the difference between calculated and observed data (thereby minimizing the target function). The parameter adjustment method is hence called „optimization algorithm". Several algorithms are available, and the choice depends on the employed target function, numerical stability and convergence radius (Rupp, 2009). The minimization takes the function values – and in some cases also gradient and curvature – into account.

5.2.2 R values

Structure factor amplitudes $|F_{calc}|$ can be calculated from a given structure model. Their disagreement with the observed structure factor amplitudes $|F_{obs}|$ is given by the crystallographic R factor, also called R value.

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

To calculate the linear residual, the two data sets need to be scaled to each other (Einspahr & Weiss, 2011).

R_{free} is the cross-validation equivalent of the crystallographic R value (Brünger, 1992). It is based on that part of the measurement data which is excluded from refinement. In this work, the so called R_{free} set were always 5% of the data. This set is chosen either randomly among reflections, or in thin resolution shells, to avoid bias between NCS-related intensities. The remaining reflections, which are used for refinement, are called working set, and the R value derived from them is called R_{work} (Rupp, 2009; Weiss, 2001).

The more data there is, the better a given number of parameters can be determined. The fewer parameters there are, the better they can be determined by a certain number of data. If there are more parameters than data, no unique solution is possible. Therefore, the data-to-parameter ratio is of crucial importance to crystallographic refinement. If the model describes random errors, because of the introduction of too many parameters, this is called overfitting: The additional parameters improve the fit between observed and calculated data beyond the experimental evidence. The R_{free} can be an indicator of overfitting (Rupp, 2009). Also a good indicator is the difference $R_{free} - R_{work}$ (ΔR) and the quotient R_{free}/R_{work}. Urzhumtseva *et al.* (2009) showed that the distribution maximum of the ΔR values is about proportional to the logarithmic resolution $ln(d)$:

$$\Delta R = \frac{2.0+2.4ln(d)}{100}$$

Tickle *et al.* (2000) derived:

$$R_{free}/R_{work} = \sqrt{\frac{N+P}{N-P}}$$

with N being the number of reflections (proportional to d^3) and P being the effective number of parameters (including both constraints and restraints). Sheldrick (personal communication) proposed the relation

$$R_{free}/R_{work} = \sqrt{\frac{1+Q}{1-Q}} \text{ with } Q = 0.025 \cdot p \cdot d^3 \cdot (1-s)$$

For p, the number of parameters per atom, 1.5 is proposed for restrained isotropic and 3.0 for restrained anisotropic refinement. s is the fractional solvent content. This equation can also be used as

$$p = \frac{(R_{free}/R_{work})^2 - 1}{(R_{free}/R_{work}) \cdot 0.025 \cdot d^3 \cdot (1-s)}$$

to get an approximation for the effective number of parameters per atom. In this chapter, this value is referred to derived number of parameters (d.n.p.), but it takes also constraints into account.

5.2.3 Restraints and constraints

To heighten the data-to-parameter ratio, restraints and constraints are applied. Both are derived from general valid observations – from our prior knowledge.

Restraints are dependencies that have not to be fulfilled exactly. They are treated like data and have a target value as well as an uncertainty within which they should be met. They are added to the target function. In SHELXL, which uses a least squares residual:

$$M = \Sigma[w_X(|F_{obs}|^2 - |F_{calc}|^2)^2] + \Sigma[w_R(T_{target} - T_{calc})^2]$$

w_X X-ray weights, in SHELXL $w_X(|F_{obs}|^2 - |F_{calc}|^2)^2 = 1$

w_R restraint weight $w_R = \sigma^{-2}$

By the SHELXL definition of w_X, w_R becomes independent from resolution and structure. Also, w_X increases if the $|F_{obs}|^2$ and $|F_{calc}|^2$agreement improves, for example in the course of the refinement process (Sheldrick, personal communication).

A typical example for a restraint is the **`FLAT`** restraint in SHELXL (Sheldrick, 2008), which tries to bring all atoms named in a plane as for example plausible in an aromatic ring system. Restraints can be unimodal, i.e. have only one target value, or be multimodal, having several. Note that multimodal conditions are more suitable for validation than as restraints.

Shift-limiting restraints, as for example jelly-body refinement in REFMAC (Murshudov, 2010; Murshudov *et al.*, 1997), can be used to dampen parameter shifts in early refinement stages. Restraints are usually weighted generally against the reflection data.

If the restraints get too much weight, the molecule is too rigid and ideal, if the reflection data gets too much weight, overfitting may easily happen and the structure might become chemically unreasonable.

Constraints are values or dependencies that have to be fulfilled exactly. They must be adhered to, and thereby, they reduce the number of parameters to refine. Constraints come as explicit constraints, which are defined by the crystallographer, and implicit constraints, which are a result of the method. A typical example is *riding hydrogen* refinement, where the position of hydrogen is solely dependent on its carrier group's position and chemical properties. For implicit constraints, a good example is the space group, which gives the symmetry to which the structure adheres to: Space group symmetry cannot be broken by the model, as it is a presupposition by the software.

5.2.4 Atomic displacement parameters

In X-ray structure analysis, the experiments average over both measurement time and molecules. Therefore the resulting structures do not contain sharp atomic positions, but a three-dimensional probability density. Deviations are a result of internal static disorder (conformational freedom), internal dynamic disorder (vibration) as well as lattice defects and other vibrations. This probability density can be described as a spherical Gaussian centered on the mean atomic position. The Gaussian width is the mean square deviation <u^2>. In macromolecular crystallography, the B factor is a common description:

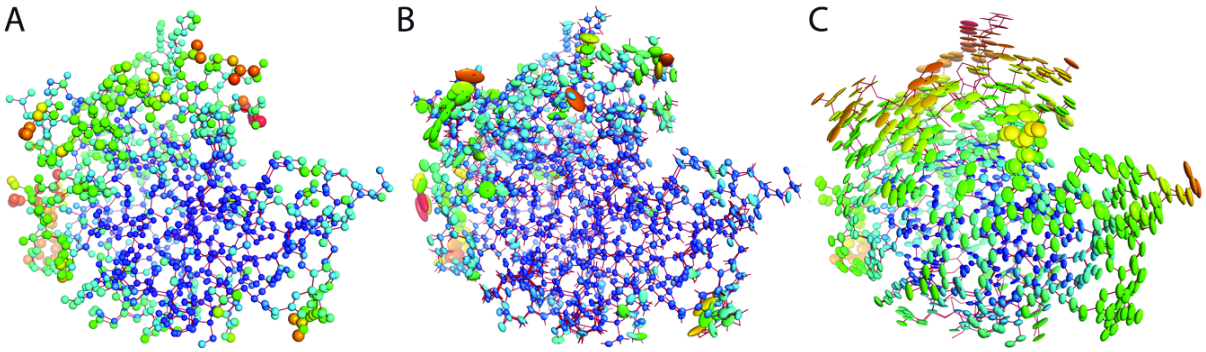

Figure 5.1: Lysozyme ADPs. **A.** Isotropic ellipsoids at 50% probability. **B.** Anisotropic ellipsoids at 50% probability. **C.** TLS; ellipsoids at 80% probability. The influence of the libration tensor is clearly visible.

$$B = 8\pi^2 < u_{iso}^2 >$$

This *isotropic* treatment results in one additional parameter for each atom.

To add detail, the probability density can be described anisotropically as a symmetric tensor with six parameters per atom (three variance and three covariance values):

$$exp(-x^T \cdot U \cdot x)$$

Here, x is the displacement vector from the equilibrium position. This is suitable for data extending to very good resolution, if the U_{ij} values are restrained, approximately up to 1.8 Å. For visualization, the tensor is often transformed into an orthogonal coordinate system. $exp(-x^T \cdot U \cdot x) = C$ describes an elliptical surface on which the probability is constant and is the basis for displacement ellipsoid plots (Johnson, 1965). If the eigenvalues of the matrix U are expressed on a suitable Cartesian basis, they correspond to the length of the prinicipal axes of the ellipsoids and the eigenvectors give the direction of these axes (Trueblood *et al.*, 1996). If the eigenvalues are negative, the atomic displacement is non-positive definite (n.p.d.). This state is physically impossible and therefore nonsense: It often hints to errors in the model or the data. For comparison, sometimes the equivalent isotropic displacement factor is calculated on orthogonal basis:

$$U_{eq} = (U_{11} + U_{22} + U_{33})/3$$

5.2.5 Established atomic displacement restraints in SHELXL

The following restraints for ADPs are established in SHELX (Schneider, 1996; Sheldrick, 2008):

```
DELU s1 s2 [atomnames]
```

The ADPs of two atoms connected directly to each other or via a 1,3-relationship are restrained in the direction of their direct connection. Different sigma can be given for 1,2- and 1,3-distances. **`DELU`** stands for "Minimizing delta U".

SIMU restrains the U_{ij} values of atoms closer than a specified maximum distance to be the similar. If they are terminal, an extra standard uncertainty can be specified for them. As **SIMU** is only a crude approximation compared with **DELU**, it should be used with looser sigma than **DELU**. The name **SIMU** stands for "Making similar U".

SIMU s st d_{max} **[atomnames]**

The restraint **ISOR** can be used to make anisotropic displacement similar in all directions within one atom. The ADP will appear more like an isotropic displacement parameter.

5.2.6 Implementation in other refinement programs

Translation, libration and screw motion of a rigid molecular domain in a crystal lattice can be described with 20 parameters, as shown by Schomaker & Trueblood (1968). This TLS decription is well below the six parameters per atom required for the free anisotropic refinement. The uniform, rigid movement of atoms can be understood as a constraint. TLS employs a strict domain definition and uses exactly 20 parameters per domain:

6 for the symmetric tensor of translation movement

6 for the symmetric tensor of libration movement

8 for the asymmetric tensor of screw motion (quadratic correlation between T and L)

The program REFMAC allows for a TLS refinement of macromolecules (Winn *et al.*, 2001) and a free refinement of an isotropic B factors per atom at the same time. However, the resulting displacement is not allowed to be non-positive definite. If an atom is n.p.d., a small number is added to the T tensor. This shows: The T tensor and the isotropic displacement are dependent on each other. In general, parameters which are refined should not be dependent on each other. While it is possible in maximum likelihood to refine them anyway, least squares refinement becomes unstable.

In PHENIX.REFINE (Adams *et al.*, 2010), the displacement of an atom consists of the three components. A symmetric tensor is calculated anisotropic effects and crystal lattice vibrations. This is applied for the each and every atom. The second term is the isotropic or anisotropic contribution for each individual atom. The third summand is called U_{group} and models the displacement resulting from concerted movements of an atomic group (Afonine, 2010). U_{group} contains in the general case a TLS contribution, an isotropic B factor for the subgroup and, depending on resolution, a term for librational movement of the side chain around a torsion bons (Stuart & Phillips, 1985). It is, however, not possible to have any atom in more than one TLS group. As a general problem, the three possible contributions to the displacement are dependent on each other.

For restraining the local displacement contribution, the term:

$$\frac{\left[U_{eq}(A) - U_{eq}(B)\right]^2}{d^{1.69}\left[\frac{U_{eq}(A)+U_{eq}(B)}{2}\right]^{1.08}}$$

for all atom combinations within d_{max} = 5 Å is minimized (Afonine, 2010).

5.2.7 The rigid-bond restraint idea

There are several problems in the TLS treatment. The domain definition proves difficult, but good efforts (see *e.g.* Painter & Merritt, 2006) have been made to facilitate the process. No atom can be in more than one domain. In loose regions, like loops, the modelling might be insufficient, especially if the domain is chosen to combine a floppy and a very rigid part of the molecule. As TLS does not allow for deviation, it can be seen as a constraint.

A flexible alternative to TLS are rigid bond restraints, first used by Rollett (1970) and implemented as **DELU** restraints in SHELXL. As the chemical bond is almost rigid, the displacement in the direction of the bond is kept similar. They hold rather well for C, N and O atoms in accurate small molecule structures (Rosenfield *et al.*, 1978). Didisheim & Schwarzenbach (1987) showed that if a sufficient number of such restraints is applied very tightly, they asymptote to the TLS description of rigid body motion. **DELU** is usually applied to 1,2- and 1,3-distances in anisotropic refinements with SHELXL. In this chapter, we investigate the extension of the rigid bond restraints to much greater distances (8 Å, 10 Å) than normally employed, so each atom is held by more than 20 such restraints, which should permit a flexible approach to the rigid TLS limit. These TLS restraints (**TLSR**, **REST**) have been incorporated into a test version of SHELXL-2018.

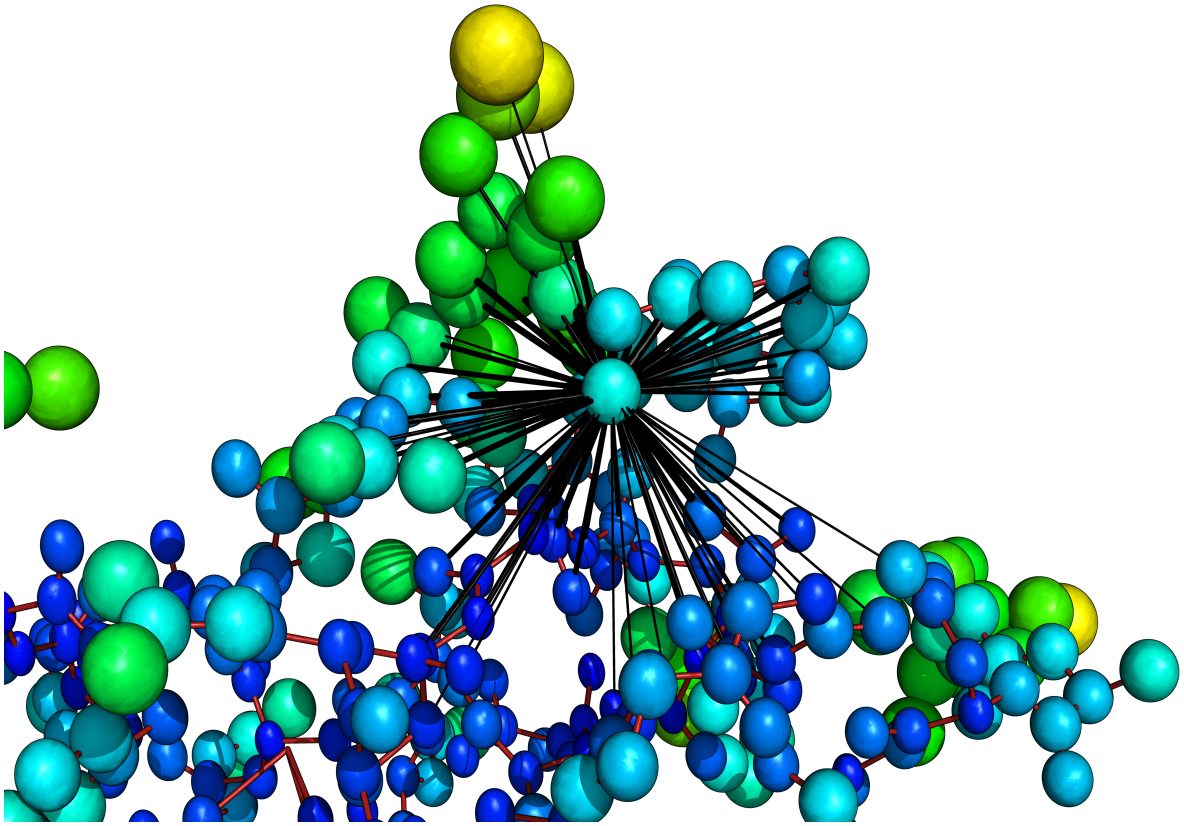

Figure 5.2: Visualization of restraint relations. The ADP of the atom in the middle is restrained to be similar to that of all neighbouring atoms within 8 Å; the line thickness stands for the strength of the applied restraint in SHELXL.

5.2.8 Implementation of the rigid-bond restraint TLSR

The command '**TLSR**' was chosen for the new rigid bond restraint. For each pair of atoms closer to each other than 8 Å, the components of their anisotropic displacement along the line joining them are restrained to be equal, with a standard uncertainty (s.u.) proportional to the square root of their distance $\mathbf{d}_{max}$ (Fig. 2). SHELXL_TLS1 contained the following implementation:

TLSR σ $\mathbf{d}_{max}$

σ If given positive, U_{ij} values along a line connecting an atom pair closer than d_{max} are restrained. If negative, $\Delta U = U_{ij} - U_{eq}$ is restrained instead of U_{ij}. One can say that only the anisotropic part is restrained.

$\mathbf{d}_{max}$ **:** The maximum distance for two atoms (at the start of the refinement) to be considered in the restraint. If the value is given positive, the weight applied is w = $\frac{1}{d \cdot \sigma^2}$; if given negative, it is $\frac{1}{d^2 \cdot \sigma^2}$.

The restraint weighting is then, depending on positive or negative σ:

$$w = \frac{1}{\sigma^2 d^p \left[|U_{ij}(A) - |U_{ij}(B)|\right]} \quad \text{or} \quad w = \frac{1}{\sigma^2 d^p \left[|U_{ij}(A) - U_{eq}(A)| + |U_{ij}(B) - |U_{eq}(B)|\right]} \quad \text{with p = 1 or 2}$$

5.2.9 Implementation of XNPD and the rigid-bond restraint REST

To be able to adjust the restraint further, **TLSR** was replaced by **REST** in the program versions SHELXL_TLS0, SHELXL_TLS2 to SHELXL_TLS9. **REST** worked on isotropic and anisotropic displacements with restraints that could be regulated separately. The command for this was:

REST σ_{iso} σ_{rest} $\mathbf{d}_{max}$ **[atom names]**

U_{eq} was assumed as the average displacement in all directions. For each atom pair A and B that was in at the beginning of the refinement nearer than d_{max}, U_{eq} was restrained to be similar. These restraints were weighted:

$$w = \frac{1}{\sigma_{iso}^2 d^p \left[|U_{eq}(A)| + |U_{eq}(B)|\right]^q}$$

Also, the displacements in the direction between the atoms A and B are restrained to be similar, in the same way as in **TLSR**. This restraint was weighted:

$$w = \frac{1}{\sigma_{rest}^2 d^p \left[|\Delta U(A)| + |\Delta U(B)|\right]^q}$$

If σ_{rest} (or σ_{iso}) were set to 0, this part of the restraint was not used. If the restraint on U_{eq} is not used (σ_{iso} = 0; command **REST 0** σ_{rest} $\mathbf{d}_{max}$ **[atom names]**), the restraint is similar to **DELU**. If both σ_{iso} and σ_{rest} are very small, so the restraint becomes tight, it should asymptote an TLS constraint, as stated by Didisheim & Schwarzenbach (1987).

After the first test series (see next section), it became clear that a constraint against non-positive ADPs was needed. The command chosen for this constraint in SHELXL was

XNPD $\mathbf{U}_{min}$

The U_{ij} values are orthogonalized and the eigenvalues are checked. If they are below a certain given cut-off value $\mathbf{U}_{min}$, they are set back to that value. PHENIX.REFINE uses a very similar method (Afonine, 2010). This proves to be an efficient and easy way to hinder ADPs from becoming too small.

5.3 Test procedures

5.3.1 Test structure preparation

	name	PDB	residues/ASU	resolution
ar66	human aldose reductase	1us0	311	0.658 Å
c2b	C2B domain of rabphilin-3A	2cm5*	154	1.192 Å
caufd	*clostridium acidurici* ferredoxin	2fdn	55	0.939 Å
cmti	squash trypsin inhibitor	1lu0	58	1.032 Å
conca	concanavalin A	**	237	1.701 Å
gico	glucose isomerase	**	386	1.542 Å
hipip	reduced high-potential iron protein mutant	1b0y	85	0.930 Å
p1lys	hen egg-white lysozyme	2vb1	129	1.100 Å
tenda	α-amylase inhibitor tendamistat	1ok0	74	0.930 Å
thox	thaumatin	1rqw	207	1.050 Å

* The deposited data has been cut at a different resolution.
** These structures are not yet deposited.

Each test structure was processed as follows: 5% of all reflections were selected with the script UNIQUEIFY (CCP4) randomly. With MTZ2HKL an **HKL** file was generated in which the same reflections are flagged as in the original **MTZ** file.

Using the program SHELXPRO (Sheldrick, 2008), the water molecules and hydrogen atoms were deleted from the structure and the displacement was set to an isotropic standard value. The occupancy of the main conformation was set to 1, and disorder, if present, removed.

Several cycles of refinement in REFMAC followed. The restraints given in the REFMAC monomer library were used. If the ligand's geometry was not present in the library, the automatically generated restraints were examined and used. Water molecules were generated with COOT (Emsley & Cowtan, 2004; Emsley *et al.*, 2010) at difference density peaks higher than 4σ, and were kept if they made chemical sense.

After refinement, the water positions were checked and edited. Whether side chains had to be „swapped" was determined by MOLPROBITY (Chen *et al.*, 2010), and eventually done. The weighting factor was determined with the automatic weighting routine in REFMAC, which does not judge by the converged negative log likelihood gain, but by the r.m.s. (bond length). After refinement, an **INS** file was generated using SHELXPRO. Restraints for non-amino acids were generated manually or by the PRODRG (van Aalten *et al.*, 1996) web service. Where applicable, **SADI** was preferred to **DFIX**, as it is less prone to systematic errors. Hydrogens were included

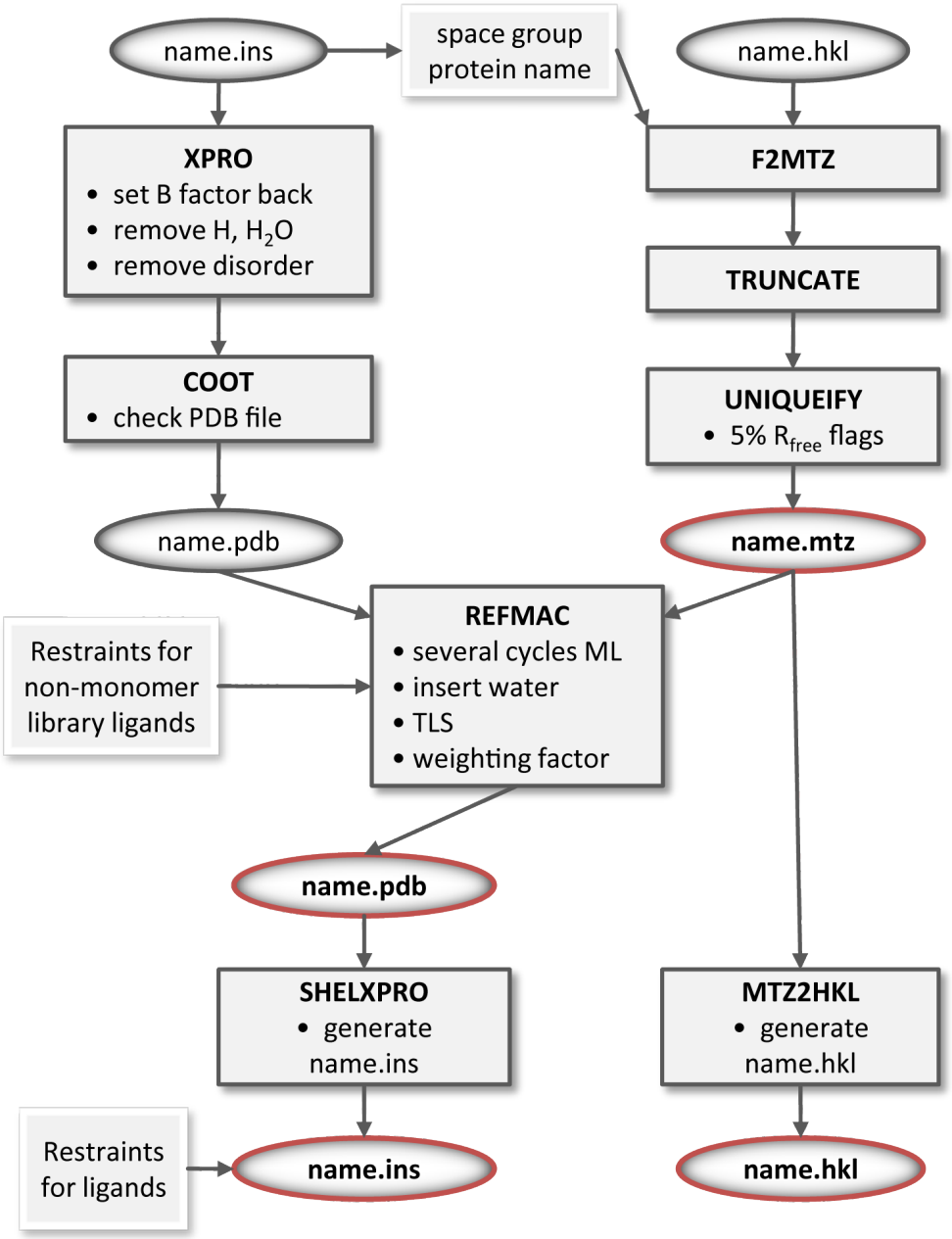

Figure 5.3: Schematic flow of the test structure preparation.

in riding positions, except at amino and hydroxy groups. This was because their protonation state could not be determined from the X-ray data.

These structures were not only used for the tests given here, but proved also useful for other projects in our lab, as they have been prepared in the same way and therefore give comparable results.

5.3.2 SHELXL-O-MATIC

To do systematic testing in a big multi-parameter space, a high-throughput script for SHELXL was constructed. This PYTHON program reads an input model in **INS** format, the data in **HKL** format and a special input file with the parameters to be varied. Also, a version of SHELXL suitable for the test is needed. The script was named SHELXL-O-MATIC.

SHELXL-O-MATIC allows central regulation of:

- parameterization of restraints
- regions of the input model on which the restraints should work
- resolution
- program version to be used

It can be used to test refinement quality against resolution and parameter ranges. The script also discards output files as specified by the user to save hard drive memory. Quality indicators from the SHELXL **LST** file as well as derived indicators are tabulated directly in the process, and are given on the screen. Within the program, several options exist to get the data plotted by GNUPLOT. The program automatically sets up a GNUPLOT script according to specifications and runs it. Axis labelling, specification from which parameter test the plot was derived etc. are automatically passed on to be shown in the plot. (A typical plot is shown in 5.4.)

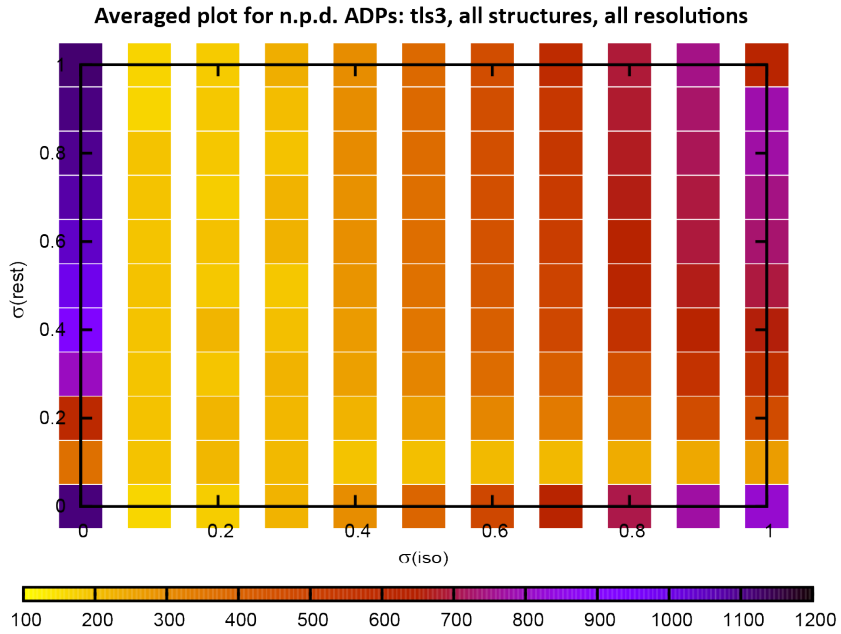

Figure 5.4: Typical output plot for test series 1. The bottom scale refers to the number of atoms n.p.d. ADPs.

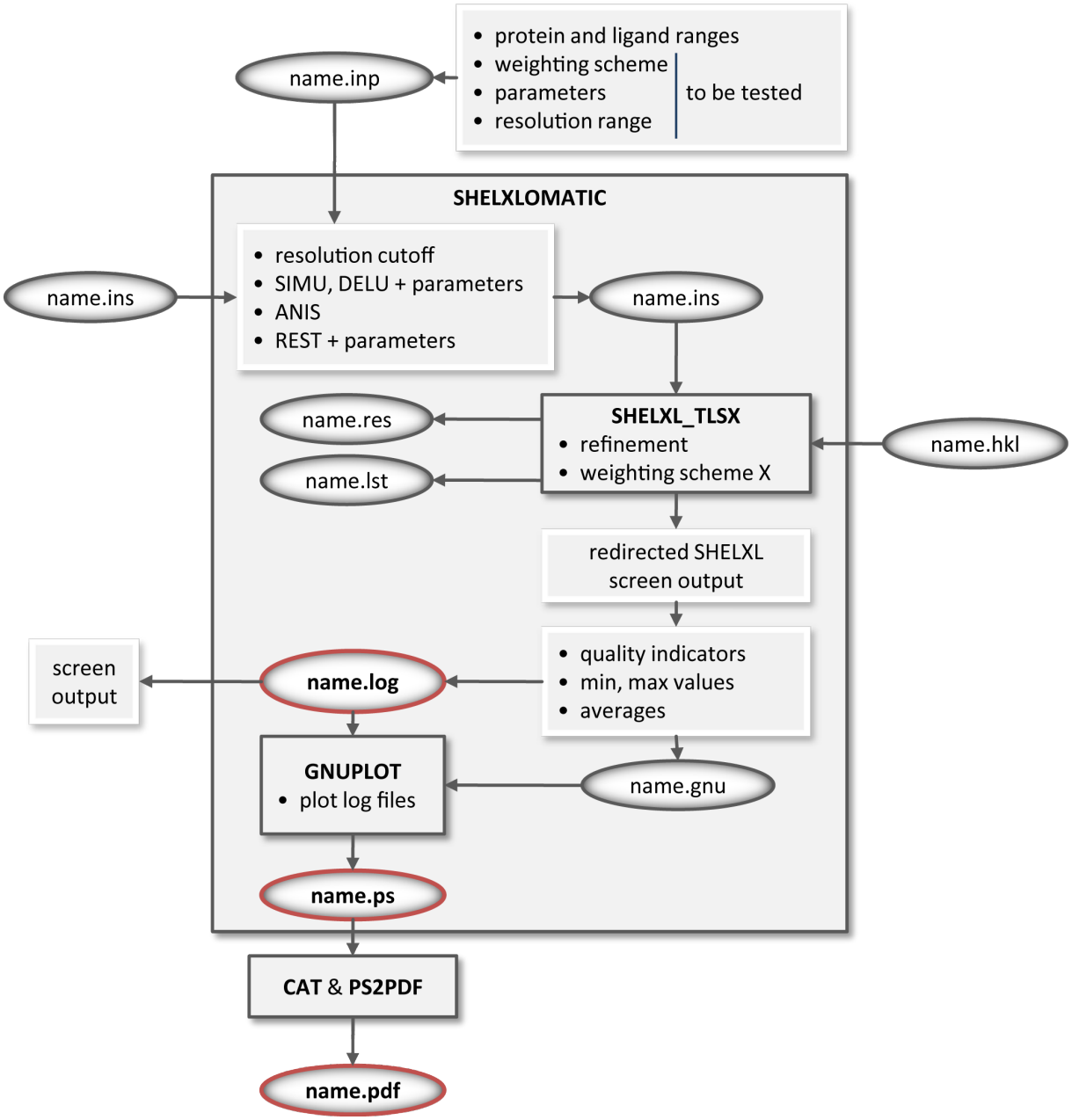

Figure 5.5: Schematic flow of the script SHELXL-O-MATIC

It showed that SHELXL-O-MATIC provides also a good benchmark test for CPUs and RAM. It clearly shows that the Intel i7 processor is superior to a hyper-threading quad core system when using multiprocessor-SHELXL. (Between the common macromolecular refinement programs, only SHELXL is capable of multiprocessor usage.) We were even able to determine by the program's performance that one processor was missing from one of the quad core workstations used. At a later stage of the project, a number of smaller PYTHON scripts were written for smaller tasks in analysis. These included logarithmic plots, average values and the program FAILFINDER which analyzes program aborts and their reasons. The high degree of automation allowed for extensive parameter tests and fast result evaluation.

Evaluated were mainly:

- R_{free}, which should be as low as possible. R_{free} should be independent of overfitting.
- $R_{free} - R_{work}$ ("ΔR") as an indicator of overfitting. It should be roughly between 10% and 4%. R_{free}/R_{work} was used instead where applicable, as its optimal value is resolution-dependent. The derived number of parameters can be calculated from this quotient.
- the number of atoms with non-positive definite ADPs ("npd")
- the number of and reasons for program aborts

5.4 Test details

In this section, all test parameters which were tested against each other are given along with the reference name of this test. In each and every refinement, solvent molecules were isotropic, and the refinement was carried out using the **`CGLS 30 -1`** command. The 'weighting scheme' refers to the SHELXL version used, as all of them weighted the restraints differently.

DELU optimization

commands used
`ANIS [protein + ligand atoms]`
`DELU [`σ**`] [protein + ligand atoms]`**
`REST 0 0.0001 10 [protein atoms]`

test set	cmti, conca, gico, hipip, thox
weighting schemes	TLS0, TLS1–9
resolutions [Å]	1.032, 1.2, 1.4, 1.6, 1.8
σ	0.0001,0.0005, 0.001, 0.005, ..., 1

Preliminary test 1

commands used
`ANIS [protein + ligand atoms]`
`DELU 0.05 [protein + ligand atoms]`
`TLSR [`σ_{rest}**`] 8 [protein atoms]`**

test set	conca, gico, thox, hipip, cmti
weighting schemes	TLS1
resolutions [Å]	1.8
σ_{rest}	-0.001, -0.01

The same refinements were carried out without REST (but with ANIS and DELU) as well as completely isotropic (but with `SIMU 0.1 [protein + ligand atoms]`).
In these tests, U_{ij}-U_{eq} was restrained instead of U_{ij}.

Preliminary test 2

commands used
ANIS [protein + ligand atoms]
DELU 0.05 [protein + ligand atoms]
TLSR [σ_{rest}] 8 [protein atoms]

test set	conca
weighting schemes	TLS1
resolutions [Å]	1.0, 1.1, ..., 2.6
σ_{rest}	-0.001, -0.01

The same refinements were carried out without rigid-bond restraint (but with **ANIS** and **DELU**) as well as completely isotropic (but with **SIMU 0.1 [protein + ligand atoms]**). In this tests, U_{ij}-U_{eq} was restrained instead of U_{ij}.

Preliminary test 3

commands used
ANIS [protein + ligand atoms]
TLSR [σ_{rest}] 8 [protein atoms]

test set	thox
weighting schemes	TLS1
resolutions [Å]	1.0, 1.1, ..., 3.5
σ_{rest}	0.001, 0.01

In these tests, U_{ij} was restrained. For comparison, a pure isotropic refinement test (with **SIMU 0.1 [protein + ligand atoms]**) was carried out.

REST test series 1

Not all test library structures were tested with all resolution because of their original resolution. Tested were: 0.8 Å (ar66 only), 1.0 Å (only ar66, caufd, hipip, tenda, caufd), 1.5 Å(all but conca and gico). σ_{iso}= 0 or σ_{rest} = 0, repectively, equals no restraint.

commands used
ANIS [protein + ligand atoms]
DELU 0.05 [protein + ligand atoms]
REST [σ_{iso}] [σ_{rest}] 10 [protein atoms]

test set	ar66, c2b, caufd, cmti, conca, gico, hipip, p1lys, tenda, thox
weighting schemes	TLS0, TLS2 – TLS9
resolutions [Å]	0.8, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5
σ_{iso}	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 0
σ_{rest}	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 0

test set	ar66, c2b, caufd, cmti, conca, gico, hipip, p1lys, tenda, thox
weighting schemes	TLS0, TLS2 – TLS9
resolutions [Å]	0.8, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5
σ_{iso}	0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1.0
σ_{rest}	0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1.0

XNPD optimization

commands used
ANIS [protein + ligand atoms]
DELU 0.05 [protein + ligand atoms]
SIMU 0.1 $C_* $N_* $O_* $S_*
XNPD [U_{min}]

test set	ar66, c2b, caufd, cmti, conca, gico, hipip, p1lys, tenda, thox
weighting schemes	TLS0
resolutions [Å]	1.5 (except gico and conca), 1.8 (gico and conca)
$\mathbf{U}_{min}$	0.001, 0.002, 0.005, 0.01, 0.02, 0.03

REST test series 2

σ_{iso} = 0 or σ_{rest} = 0, repectively, equals no restraint.

commands used
ANIS [protein + ligand atoms]
DELU 0.05 [protein + ligand atoms]
REST [σ_{iso}] [σ_{rest}] 10 [protein atoms]
XNPD 0.0020

test set	ar66, c2b, caufd, cmti, conca, gico, hipip, p1lys, tenda, thox
weighting schemes	TLS0, TLS2 – TLS9
resolutions [Å]	native resolution of each test structure, 2.0
σ_{iso}	0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1.0, 0 (equals no restraint)
σ_{rest}	0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1.0, 0 (equals no restraint)

The test structures conca and gico were not tested at their native resolutions, as they were near to 2.0 Å.

5.5 Test results

5.5.1 SIMU

Both **SIMU** as well as **DELU** are routinely used in macromolecular refinement with SHELXL. Their impact and combination potential with the new restraint were evaluated preliminary. **SIMU** makes the displacement of atoms within a certain specified radius more equal by restraining each U_{ij} value to the one of the neighbour atom. In combination with the **TLSR** restraint it was found that R factors became much higher. **SIMU** was omitted for all further tests.

5.5.2 DELU optimization

The **DELU** command restrains the displacement only in the direction of of 1,2- and 1,3-distances between atoms. This is a rigid bond restraint and already quite similar to the **TLSR** routine. Generally, the difference between R_{free} and R_{work} decreases with better resolution and with tighter **DELU** restraints.

DELU was tested in combination with **TLSR**, and later with **REST**, but the weighting scheme did not have much influence on the test outcome. The optimum value is approximately **DELU 0.05** which was chosen as a fixed value for all further tests.

DELU supplements **TLSR** well, making the displacement in the direction of bonds even more similar than for non-bonded atom pairs. Also, while **DELU** works on the full U_{ij} value, **TLSR** could work on U_{ij}-U_{eq}. This might be an advantage of the method.

5.5.3 Preliminary tests

With the **TLSR** restraints three preliminary tests were attempted.

Comparison at 1.8 Å resolution Five structures from the test library were refined isotropic, anisotropic and anisotropic with **TLSR** restraints at 1.8 Å resolution. R_{free}/R_{work} should be (by the formula given in 5.2.2 on page 60) 1.069 for isotropic refinement (appr. 1.5 parameters/atom) and 1.120 for anisotropic refinement (appr. 4 parameters/atom), assuming an average protein crystal solvent content of 0.45.

Table 5.1: Preliminary comparison between anisotropic, isotropic and anisotropic refinement with the new restraints.

	R_{free}			R_{free}/R_{work}			n.p.d.		
	iso	**tlsr**	**anis**	**iso**	**tlsr**	**anis**	**iso**	**tlsr**	**anis**
cmti	25.2%	23.3%	25.3%	1.454	1.361	1.829	7	0	17
conca	30.0%	26.8%	29.9%	1.283	1.234	1.633	5	0	219
gico	26.8%	26.1%	27.0%	1.365	1.314	1.690	105	20	1424
hipip	22.7%	21.0%	22.1%	1.513	1.543	1.881	0	2	88
thox	20.8%	20.5%	20.6%	1.313	1.298	1.530	0	0	117

Both R_{free} and the number of non-positive definite displacement parameters (n.p.d.) is lower if **TLSR** restraints are used, compared with isotropic and anisotropic refinement. But the R_{free}/R_{work} value is high in all tests, indicating more parameters being fitted than estimated above. As lowest values can be seen for refinement with **TLSR**, this gave a hint that the restraints heightened the data-to-parameter ratio.

Concanavalin A Concanavalin A was refined in a resolution range from 1.0 Å to 2.6 Å with rigid-bond restraints at σ_{iso} = 0.01 and 0.001. The **TLSR** restraints were set to working only on U_{ij}-U_{eq} and the results were compared to normal isotropic.

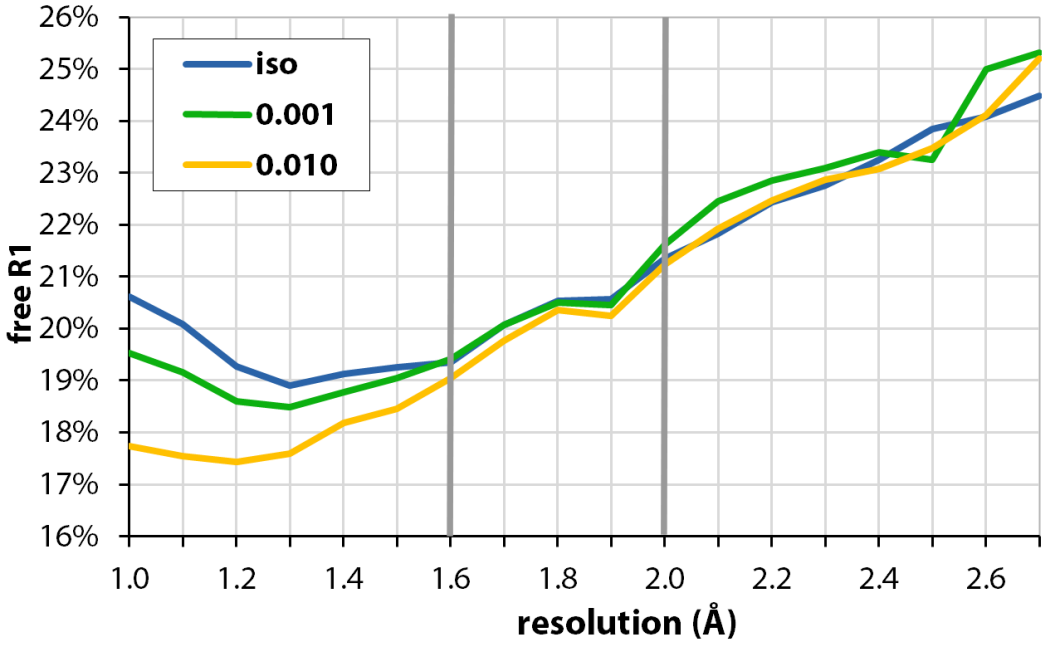

Figure 5.6: Refinement of concanavalin A with isotropic displacement parameters (iso) and with **REST** at two different standard uncertainties (0.01 and 0.001).

The **TLSR** restraint gives an equal or lower free R value than classical restrained isotropic refinement over a wide resolution range. Fig. 5.6 shows that up to a resolution of 1.6 Å, the weaker **TLSR** restraint gave the best R_{free} value and over 2.0 Å the tighter **TLSR** restraint performs better.

Thaumatin Thaumatin was tested in an even broader resolution range to find the working limits (see Fig. 5.7). To give less freedom to the model parameters, **TLSR** was set up here to work on the total of U_{ij}. No **DELU** restraints were applied.

It becomes clear from Fig. 5.7 that for low resolution, the restraints can be applied without subtracting the equivalent isotropic displacement parameters. For very tight restraints of this type (low s.u.) the effective data-to-parameter ratio should be improved and asymptote 20 displacement parameters. Hence, no additional **DELU** restraints are required.

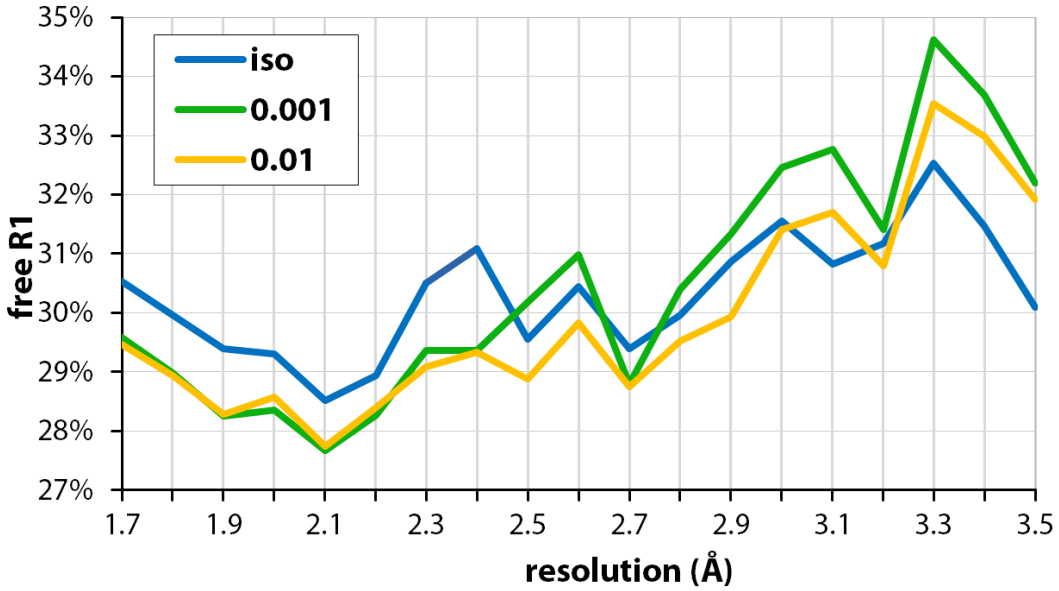

Figure 5.7: Refinement of thaumatin with isotropic displacement parameters (iso) and with **REST** at different standard uncertainties.

5.5.4 Test series 1

In the test series 1, a wide range of parameters was varied, namely, the resolution from 0.8 to 3.5 Å and the values for σ_{iso} and σ_{rest} in a broad range. Initially, the aim of the test was to find the best weighting scheme and the best values for σ_{iso} and σ_{rest} in the **REST** restraint, but as shown here, this was not possible.

Roughly 22000 refinements (with 30 cycles each) were carried out. This not only demanded a high-performance computer cluster, but also a well though-out logistics system. This holds especially true as every refinement later had to be reproducible and the distribution among computers was not implemented automatically. The two tests took about 400 hours on eight multiprocessor workstations running under SUSE or DEBIAN LINUX.

The results of the tests were examined by structure, resolution and weighting scheme against $\sigma_{iso}/\sigma_{rest}$ combinations. An overview is given in Table 5.2. R_{free}/R_{work} was not used in the evaluation, as tests were assessed over a broad resolution range.

Without the restraint (**REST 0 0 d**$_{max}$), non-positive definite ADPs and free R value became very high; the refinement became instable. At very good resolutions (0.8 – 1.5 Å), no or very weak restraints was a good choice. This was to be expected, as the high number of data at these resolutions permits also for anisotropic refinement without **REST**.

The weighting schemes TLS2, TLS3 and TLS4 gave the best results. Note that TLS3 is most similar to the weighting scheme employed by phenix.refine.

A general problem in these tests was the high number of non-positive definite ADPs and program aborts among the refinements. These obscured the results, which almost always were averages: Aborted refinements were not taken into account; and structures with many non-positive definite displacement parameters could allow the R values to be lower by modelling the physically impossible. Also, non-positive ADPs could spread by using the REST restraints

Table 5.2: Average test results for test series 1. (* Minimum R_{free} averaged over all test structures and resolutions for one combination of σ_{iso} and σ_{rest}.)

weighting scheme $w = \frac{1}{\sigma^2_{rest/iso} d^p [\lvert\Delta U(A)\rvert + \lvert\Delta U(B)\rvert]^q}$		<ΔR>	<R_{free}>	<NPD>	min. R_{free}*	aborts
tls0:	p = 0, q =0	7.8%	26.9%	49	24.08%	32
tls2:	p = 2, q = 2	7.9%	27.0%	74	**23.44%**	19
tls3:	p = 2, q = 1	8.1%	26.8%	76	23.58%	0
tls4:	p = 2, q = 0	8.3%	**26.6%**	123	24.08%	3
tls5:	p = 1, q = 2	7.8 %	27.1%	53	24.08%	50
tls6:	p = 1, q = 1	7.9%	27.0%	49	23.65%	19
tls7:	p = 1, q = 0	8.1%	26.8%	83	23.92%	8
tls8:	p = 0, q = 2	**7.6%**	27.0%	**37**	24.55%	131
tls9:	p = 0, q = 1	7.7%	27.0%	41	24.48%	68

to regions of atoms which were poorly defined in the electron density, but close to each other.

Consequently, a script named FAILFINDER to analyze program aborts was written. It could be shown that the risk of aborting increased with structure size and lower resolution. Aborts seemed not to correlate with the number of non-positive definite ADPs.

resolution	1.0	1.5	2.0	2.5	3.0	3.5
number of program aborts	2	14	52	49	81	136

Most refinements (291 of 314) aborted because the connectivity shifted and became unsuitable for the AFIX commands given to introduce riding hydrogens. This '*Bad AFIX connectivity*' abort of SHELXL is common in macromolecular refinements and could be avoided by giving a warning in the program output without stopping the refinement. The number of such errors could serve as a quality indicator, but one bad atom position shift would not halt the overall refinement. 22 refinements aborted with '*refinement unstable*'.

5.5.5 Implementation and optimization of XNPD

The constraint functionality was tested before optimization, and eventually debugged. One problem that could not be resolved within this project are the ADPs of atoms lying on special position: As they have already constrained U_{ij} values, the **XNPD** constraint can lead to contradiction. This was resolved by excluding these atoms from **XNPD** automatically.

The new constraint was optimized. As even in biological macromolecules, U_{ij} contributions can be small, too high values might give inaccurate models, despite the R_{free} improved in our tests averaged for all structures. Too small values (< 0.001) seem unreasonable for protein crystals with their highly flexible compounds. Generally, low variation occured, as shown in 5.6 on page 88. So a cut-off value of 0.002 was used as a compromise in all further tests.

5.5.6 Test series 2

With the **XNPD** constraint in operation, the **REST** restraint was again tested.

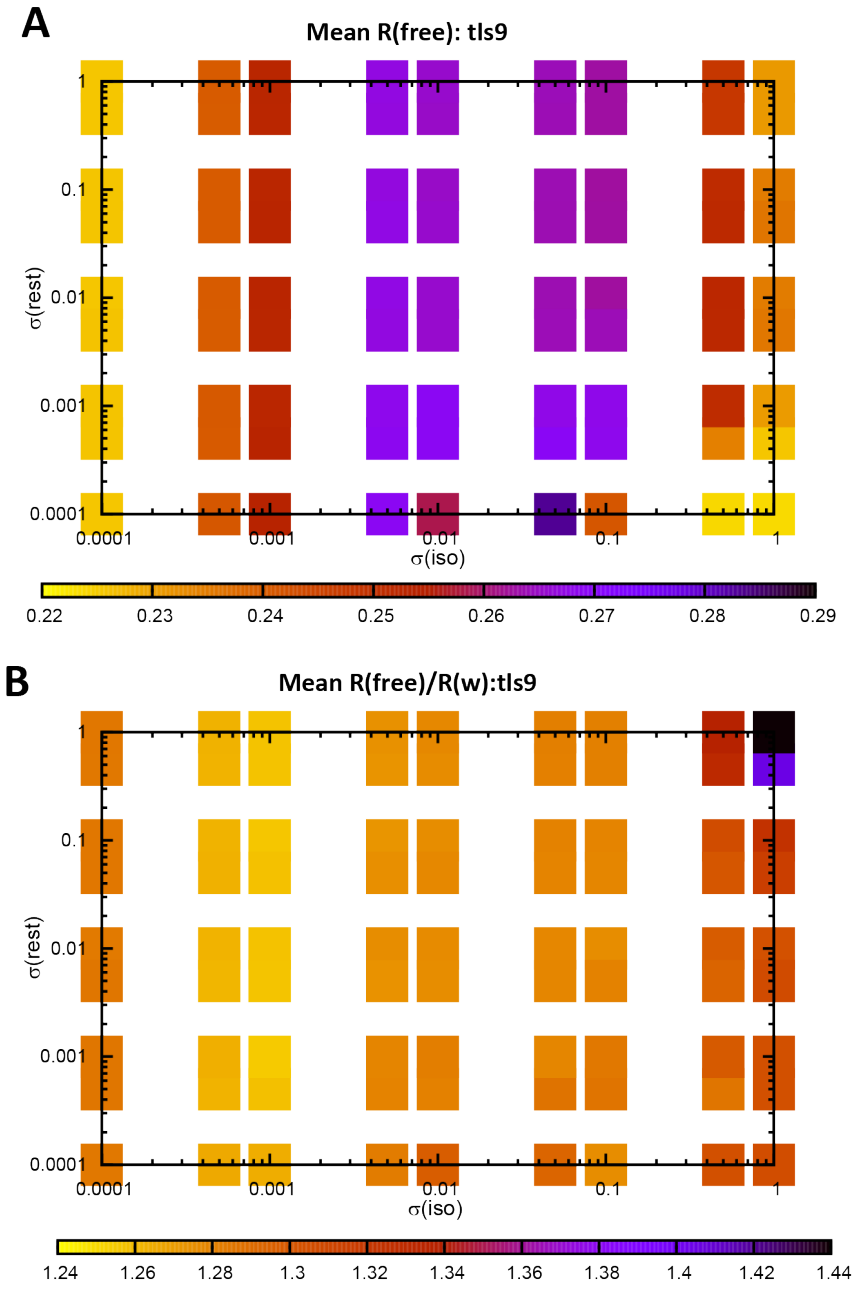

Figure 5.8: Averaged test indicators for weighting scheme TLS9 at 2.0 Å. The best <R_{free}> results with $\sigma_{iso} = 1$, $\sigma_{rest} = 0.0001$, while <R_{free}/R_{work} > is acceptable.

Table 5.3: Rigid-bond restraint **REST** in combination with **XNPD** at native resolution. The results are averaged for all tested structures. (*derived number of parameters assuming a solvent content of 45% and the average test structure resolution of 1.0 Å.)

native test structure resolution (see table 5.3.1 on page 66)					
$w = \frac{1}{\sigma^2_{rest/iso} d^p [\lvert\Delta U(A)\rvert + \lvert\Delta U(B)\rvert]^q}$	**weighting scheme**	**<R_{free}/R_{work}>**	**dnp***	**<R_{free}>**	**aborts**
tls0:	p = 0, q =0	1.1186	8.1	21.74%	0
tls2:	p = 2, q = 2	1.1231	8.4	21.44%	0
tls3:	p = 2, q = 1	1.1327	9.0	20.95%	0
tls4:	p = 2, q = 0	1.1417	9.6	20.56%	0
tls5:	p = 1, q = 2	1.1126	7.7	22.22%	1
tls6:	p = 1, q = 1	1.1216	8.3	21.57%	0
tls7:	p = 1, q = 0	1.1308	8.9	21.06%	0
tls8:	p = 0, q = 2	1.1045	7.2	22.79%	5
tls9:	p = 0, q = 1	1.1105	7.6	22.29%	1
	no REST	1.2427	15.6	17.80%	0

At native resolution, 1.37% of the refinements aborted with "*Bad AFIX connectivity*". This is a great improvement if compared to test series 1, and shows how the **XNPD** constraint stabilizes the refinement. The more freedom the refinement has (indicated by the derived number of parameters, compare Table 5.3), the lower the mean R_{free}. This shows that the **REST**-restrained model is not in good agreement with the measured structure.

The same test was repeated at 2.0 Å.

Table 5.4: Rigid-bond restraint **REST** in combination with **XNPD** at 2.0 Å resolution. The results are averaged for all tested structures. (*derived number of parameters assuming a solvent content of 45%.)

2.0 Å					
$w = \frac{1}{\sigma^2_{rest/iso} d^p [\lvert\Delta U(A)\rvert + \lvert\Delta U(B)\rvert]^q}$	**weighting scheme**	**<R_{free}/R_{work}>**	**dnp***	**<R_{free}>**	**aborts**
tls0:	p = 0, q =0	1.3031	2.4	24.51%	0
tls2:	p = 2, q = 2	1.3130	2.4	24.59%	0
tls3:	p = 2, q = 1	1.3350	2.6	24.49%	0
tls4:	p = 2, q = 0	1.3597	2.7	24.33%	0
tls5:	p = 1, q = 2	1.2912	2.3	24.59%	0
tls6:	p = 1, q = 1	1.3089	2.4	24.55%	0
tls7:	p = 1, q = 0	1.3301	2.5	24.51%	0
tls8:	p = 0, q = 2	1.2797	2.2	24.64%	0
tls9:	p = 0, q = 1	1.2878	2.3	24.53%	0
	no REST	1.6762	4.3	19.29%	0

Here, the lowest mean R_{free} is produced by the weighting scheme TLS4, while TLS8 has the lowest derived number of parameters. The results indicate that at 2.0 Å resolution, the restraint might be useful. As shown in Fig. 5.8 on page 77, again, a tight restraint on the ΔU values and a loose restraint on U_{eq} works good, balancing R_{free} and the number of parameters.

While no **REST** restraint (meaning anisotropic refinement with only **DELU**) does most definitely result in overfitting at 2 Å resolution, the values are given in Table 5.4 to get an idea how much **REST** lowers the derived number of parameters.

The structure of squash trypsin inhibitor (cmti) was chosen to compare the outcome of this test series:

Table 5.5: Comparison between ADP treatments for squash trypsin inhibitor. (* dnp calculated with 40% solvent content in squash trypsin inhibitor.)

	$\mathbf{R}_{free}/\mathbf{R}_{work}$	**dnp***	$\mathbf{R}_{free}$
isotropic in REFMAC at 1.0 Å	1.1158	7.3	23.31%
TLS in REFMAC at 1.0 Å	1.1039	6.6	20.69%
with REST at 1.0 Å	1.3040	17.3	18.49%
anisotropic at 1.0 Å	1.3166	17.9	18.42%
isotropic in REFMAC at 2.0 Å	1.5020	3.2	23.57%
TLS in REFMAC at 2.0 Å	1.4399	2.9	21.92%
with REST at 2.0 Å	1.4590	3.0	23.64%
anisotropic at 2.0 Å	1.7535	4.2	24.97%

From the table, it becomes clear that at 1.0 Å, the weak REST restraints are similar to pure anisotropic refinement, while at 2.0 Å, they model the structure (see 5.10 on page 81) very similar to the anisotropic ADPs, but with fewer parameters. However, despite TLS refinement lets the ADP ellipsoids look too "round", REFMAC achieves a much better R_{free} value with a comparable derived number of parameters. This is because other than SHELXL, REFMAC is optimized in all aspects for macromolecular refinement at medium resolutions, using a better solvent model and more specialized restraints. The aim of this project was to make SHELXL more capable of such refinements, but the new restraints are only one step in this.

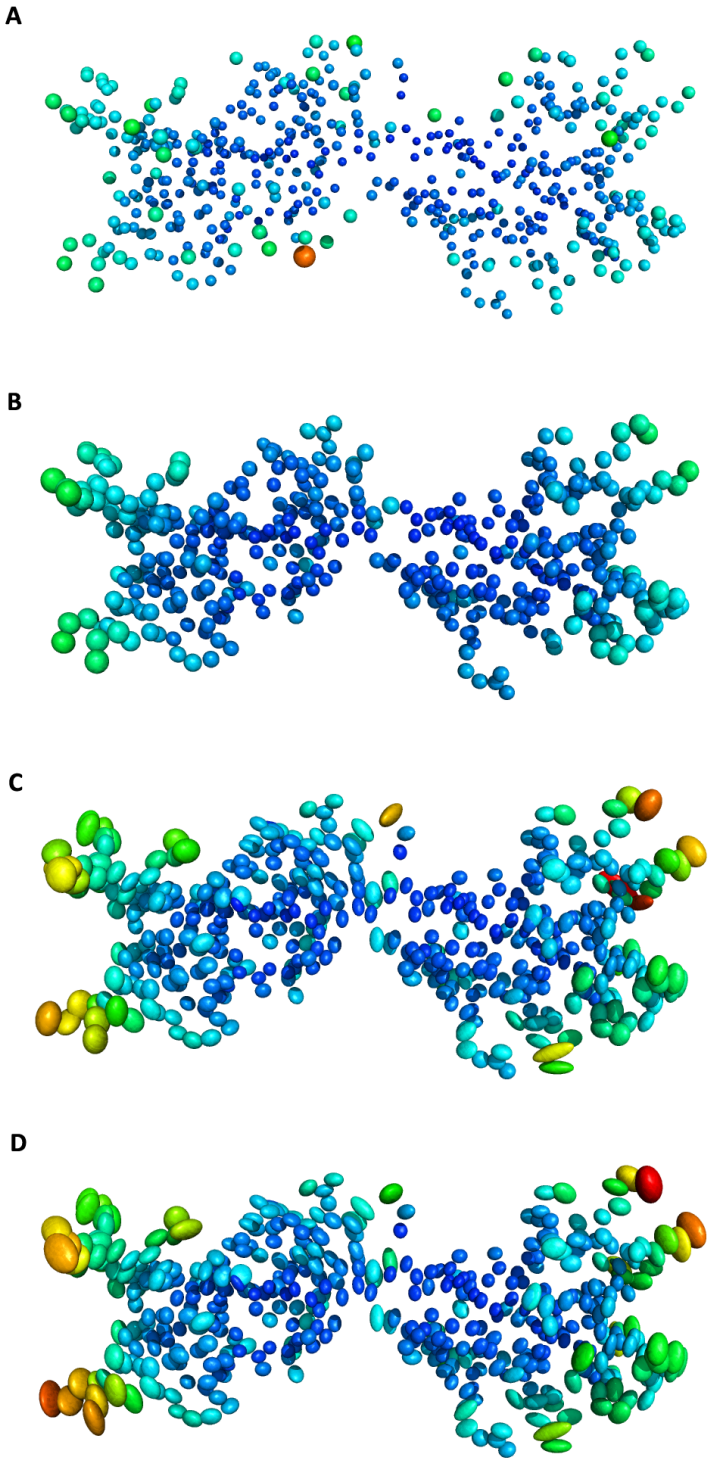

Figure 5.9: The structure of squash trypsin inhibitor refined at 1.0 Å: **A.** With isotropic ADPs in REFMAC. **B.** With TLS plus individual isotropic ADPs in REFMAC. Two domains, one for each molecule in the ASU, were used. Note that the surface displacement is underestimated, and the anisotropic contribution relatively small. **C.** Refined with SHELXL using the weighting scheme TLS6 and `REST 1 0.1`. This is almost as anisotropic refinement. **D.** With anisotropic ADPs in SHELXL. This model is the most realistic one, as it had the most freedom at a good data/parameter ratio.

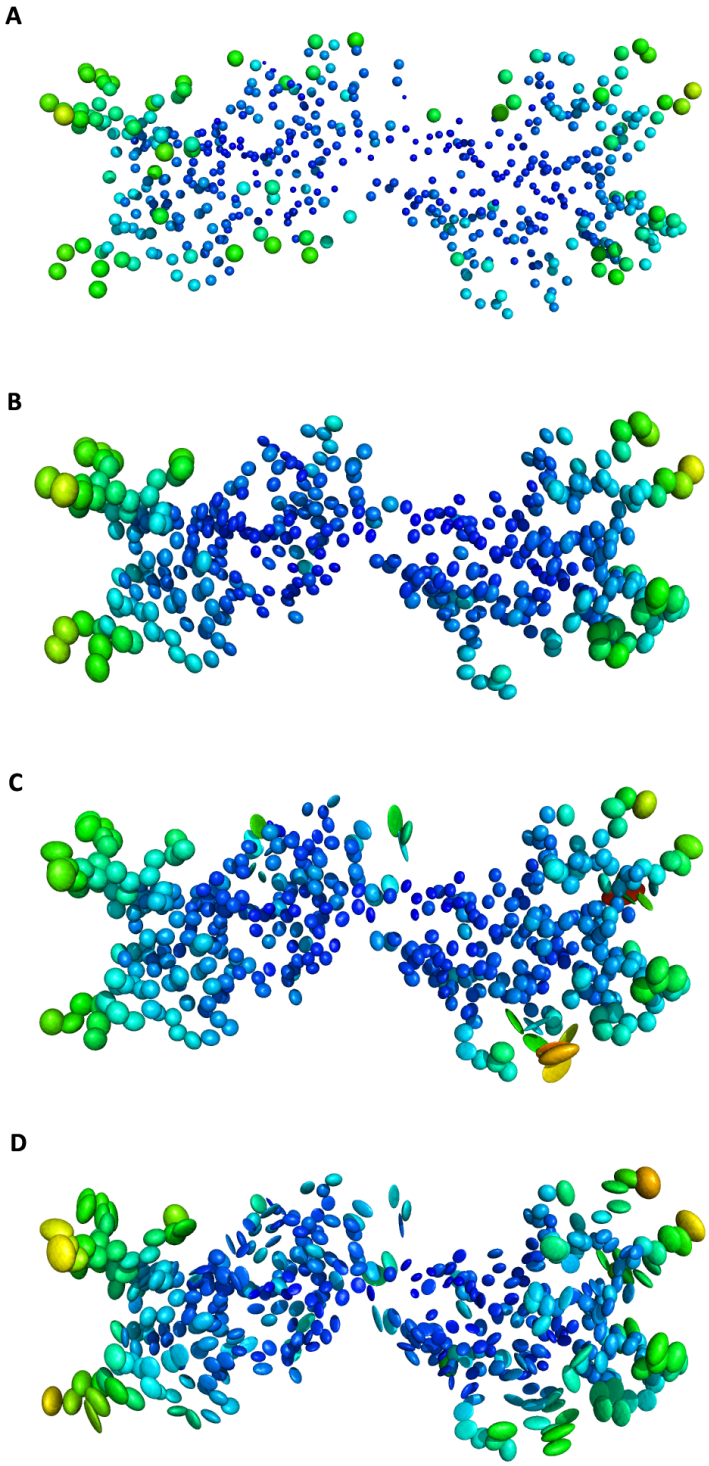

Figure 5.10: The structure of squash trypsin inhibitor refined at 2.0 Å: **A.** With isotropic ADPs in REFMAC. **B.** With TLS plus individual isotropic ADPs in REFMAC. Two domains, one for each molecule in the ASU, were used. Note that the surface displacement is underestimated, and the anisotropic contribution relatively small. **C.** Refined with SHELXL using the weighting scheme TLS6 and `REST 1 0.1` **D.** With anisotropic ADPs in SHELXL. The flat ellipsoids are results of overfitting, as anisotropic refinement needs too many parameters at this resolution. However, it is very similar to C., which needed much less parameters and was a stable refinement. The ADPs refined with `REST` restraints are too "round", but comparable to TLS.

5.6 Discussion and outlook

At medium resolution, the new **REST** restraint makes restrained anisotropic refinement with SHELX more feasible. Supplemented by **DELU**, with a high freedom in the individual isotropic contribution of atoms, and a tighter restraint on ΔU, it poses a good addition to SHELXL for macromolecular refinement at medium resolution.

A new constraint has been introduced to SHELXL, which can be used to prohibit ADPs from becoming non-positive definite: **XNPD**. **REST** restraints have to be combined with **XNPD** to avoid refinement instability.

It has been confirmed in a great number of cases (see Winn *et al.* (2001); Afonine (2010) and the two protein structure refinements described in this work) that the REFMAC treatment of an isotropic atomic displacement in combination with 20 TLS parameters lowers R_{free} significantly (0.5–2.5%). It remains obscure why TLS poses such an improvement to the refinement models of macromolecules, as its theoretical basis is not fully applicable to macromolecules.

The rigid-bond restraints have some advantages over TLS: In the more flexible or peripheral parts of a macromolecule, there are fewer near neighbors. Thus atoms are allowed to exhibit higher and more variable anisotropic motion. In contrast to the use of TLS constraints, it is not necessary to define (semi-)rigid domains, they appear naturally. Despite these advantages, REST needs more parameters than TLS, especially if not used with tight σ values.

SHELXL still performs not as good as REFMAC at medium resolution and overfitting occurs more easily (indicated by higher R_{free}/R_{work} values). REFMAC is more robust at medium to low resolution because of a more sophisticated solvent model, torsion angle and other multimodal restraints as well as a maximum likelihood target function. More facilities to make SHELXL suitable for macromolecular refinement are needed.

What becomes very clear from figures like 5.9 on page 80 and 5.10 on the previous page is that surface motion is much bigger than would be expected by TLS or from a rigid body. This motion is almost always orthogonal to the centre of the macromolecule. This might be a hint for further improvement in the modelling of atomic displacement parameters.

Appendix

Data quality indicators

Definitions for the data quality indicators discussed in this work are given here. The article by Einspahr & Weiss (2011) served as a general guideline. For calculation of all R-factors in XPREP, negative observed intensities are set to 0.

R_{int} is also called R_{merge} and in XDS (Kabsch, 2010) R-factor (observed). It gives the variation of the individual intensity measurement $I_i(hkl)$ around the average intensity $< I_{hkl} >$. For each reflection hkl, there are i observations. R_{int} is dependent on data multiplicity, which limits its indication of data quality (Weiss & Hilgenfeld, 1997).

$$R_{int} = \frac{\sum_{hkl} \sum_i |I_i(hkl) - < I_{hkl} > |}{\sum_{hkl} \sum_i I_i(hkl)}$$

A better choice is the multiplicity independent merging R-factor R_{meas}, also called R_{rim} (Diederichs & Karplus, 1997; Weiss & Hilgenfeld, 1997; Weiss, 2001). It gives the measurement precision independent of the data set's multiplicity. N(hkl) is the number of individual measurements of the reflection hkl.

$$R_{rim} = \sum_{hkl} \sqrt{\frac{N(hkl)}{N(hkl) - 1}} \cdot \frac{\sum_i |I_i(hkl) - < I_{hkl} > |}{\sum_{hkl} \sum_i I_i(hkl)}$$

The precision of averaged intensity measurements is given by the precision-indicating merging R-factor R_{pim} (Weiss, 2001).

$$R_{pim} = \sum_{hkl} \sqrt{\frac{1}{N(hkl) - 1}} \cdot \frac{\sum_i |I_i(hkl) - < I_{hkl} > |}{\sum_{hkl} \sum_i I_i(hkl)}$$

To compare two data sets, the correlation coefficient can be used. It is defined by:

$$CC = \frac{\sum_{hkl} [(I_{hkl}(A) - < I_{hkl}(A) >) \cdot (I_{hkl}(B) - < I_{hkl}(B) >)]}{\sqrt{\sum_{hkl} (I_{hkl}(A) - < I_{hkl}(A) >)^2 \cdot \sum_{hkl} (I_{hkl}(B) - < I_{hkl}(B) >)^2}}$$

Several indicators exist for anomalous data:

For the calculation of the anomalous R-factor R_{anom}, $< I(hkl) >= 0.5 \cdot (I(hkl) + I(-h-k-l)$.

$$R_{anom} = \frac{\sum_{hkl} |I(hkl) - I(-h-k-l)|}{\sum_{hkl} < I(hkl) >}$$

The ratio R_{anom}/R_{pim} is a general preliminary indicator for the strength of the anomalous signal for SAD phasing; it should be bigger than 1.5 (Panjikar & Tucker, 2002; Rupp, 2009).

$d''/\sigma(d'')$ is the anomalous signal-to-noise ratio. Its average in resolution shells is often calculated and plotted to find the right cut-off in resolution. A value of $\sqrt{2/\pi} \approx 0.8$ is considered to be noise and convergence to this value in the outer shells can indicate proper data processing. The term d'' is $|I_{hkl}\text{-}I_{-h-k-l}|$ and $\sigma(d'')$ is the corresponding estimated standard uncertainty for this anomalous difference.

The dependence between anomalous differences of two given data sets is indicated by the anomalous correlation coefficient CC_{anom}. If only one data set is available, the $CC_{anom,self}$, can be calculated by randomly partitioning a data set into two.

$$CC_{anom} = \frac{\sum_{hkl} [(\Delta I_A - < \Delta I_A >) \cdot (\Delta I_B - < \Delta I_B >)]}{\sqrt{\sum_{hkl} (\Delta I_A - < \Delta I_A >)^2 \cdot \sum_{hkl} (\Delta I_B - < \Delta I_B >)^2}}$$

ΔI are the anomalous differences $I(hkl) - I(-h-k-l)$ of the compared data A and B. A CC_{anom} value above 30% signifies a moderate anomalous signal.

Graphics software

In this work, the following programs were used to generate figures:

software	usage	manufacturer
Adobe CS 4	graphics editing	Adobe systems, San Jose, USA
CCP4	map conversion	Collaborative Computational Project (1994)
Chem Draw Pro 9	2D molecule figures	Cambridge Soft, Cambridge, USA
GNUPLOT 4.0	mathematical plots	Thomas Williams, Colin Kelley
HKL2MAP	quality indicator plots	Pape & Schneider (2004)
Open Office Calc	data preparation for plots	Open Office community
Excel 2010	mathematical plots	Microsoft
PYMOL	molecule rendering	DeLano Scientific, Palo Alto, USA
XPREP	quality indicator plots	Bruker AXS, Madison, USA

Multi-solution approach

PHASER/SHELXE tries for Concanavalin A.
The calculation number refers to the archive in our lab.

No.	PHASER			SHELXE −a30 XX.pda								model
				-y1.6		-y2.0		-y2.5		-y3.0		
no	RFZ	TFZ	LLG	CC	AA	CC	AA	CC	AA	CC	AA	
2	4.1	5.4	21	5.68%	13.67	8.11%	11	8.46%	11.2	7.24%	13.3	1qmo_cut1
4	3.4	4.4	14	7.50%	11.4	4.15%	17.5	9.48%	10.8	7.51%	11.2	2b7y_mod
8	4.8	8.8	73	9.80%	17.5	**33.56%**	**26**	6.49%	12.8	6.34%	13.3	2ltn_cut3
9	5.8	6.8	119	**31.51%**	**23.88**	**35.67%**	**29**	8.76%	10.8	7.86%	8.33	2ltn_cut4
10	5.1	6.3	86	6.74%	10.75	11.07%	16.5	8.62%	12.7	9.33%	12.6	2ltn_cut5
11	5.0	4.5	71	7.80%	14	5.63%	8.833	8.69%	12	7.33%	13.5	2ltn_cut6
12	4.5	4.8	52	7.75%	16	5.80%	9.75	8.04%	9.17	4.83%	10.8	2ltn_cut8
13	4.6	4.4	42	5.93%	8.5	7.35%	9.4	7.24%	13	9.25%	15.5	2ltn_cut7
14	4.2	5.0	38	9.44%	10.86	9.76%	15.5	6.24%	10.6	4.59%	7.75	2ltn_cut9
15	6.1	4.1	39	7.38%	9.8	6.77%	12.2	9.14%	12.3	8.03%	11.6	1g7y_cut1
17	4.6	4.8	35	6.46%	14.25	**34.36%**	**19.36**	8.01%	9.33	7.23%	11.8	1ioa_cut1
19	4.1	4.9	25	8.36%	10.57	9.98%	10.17	7.16%	12.8	9.62%	9.88	1lgc_mod
20	4.7	4.9	25	4.82%	12.67	4.86%	13.5	8.55%	19.3	7.06%	12.5	1loa_cut1
22	4.8	5.3	37	10.18%	9.857	6.35%	12.67	5.84%	9.75	6.39%	9.6	1qmo_cut2
23	4.4	4.1	23	6.42%	8	11.39%	15.75	9.16%	14.5	5.95%	14	2ltr_cut1
24	4.2	4.3	25	6.39%	10.6	9.48%	10.57	5.34%	10	9.00%	12.8	2ltr_mod1

XNPD test results

XNPD test. No big variation occurs in the test described in section 5.4 on page 72. For the R_{free}/R_{work} target values, an average solvent content of 0.45 as well as 1.5 parameters per atom for isotropic refinement and 4.0 for anisotropic refinement were assumed.

	R_{free}/R_{work} against resolution			R_{free} values against resolution		
XNPD <U$_{min}$>	**1.50 Å**	**1.80 Å**	**2.10 Å**	**1.50 Å**	**1.80 Å**	**2.10 Å**
0.001	1.451	1.603	1.714	19.16%	22.50%	23.46%
0.002	1.450	1.604	1.718	19.15%	22.52%	23.50%
0.003	1.451	1.603	1.721	19.16%	22.51%	23.53%
0.005	1.450	1.604	1.716	19.15%	22.53%	23.47%
0.010	1.448	1.598	1.715	19.12%	22.46%	23.50%
0.020	1.444	1.591	1.707	19.08%	22.40%	23.44%
0.0300	1.440	1.580	1.697	19.05%	22.33%	23.38%
Target value iso:	**1.072**	**1.128**	**1.213**			
Target value anis:	**1.206**	**1.394**	**1.754**			

Abbreviations

5'UMP	uridine monophosphate
AA	average chain length in SHELXE
ADP	atomic displacement parameter
ASU	asymmetrical unit
CC	correlation coefficient
cDNA	complementary DNA
dnp	derived number of parameters
ER	endoplasmic reticulum
et al.	*et alii*
Fig.	Figure
LLG	log likelihood gain
MAD	multiple wavelength anomalous diffraction
MIR	multiple isomorphous replacement
MR	molecular replacement
NCS	non-crystallographic symmetry
npd	non-positive definite (also used: n.p.d.)
RFZ	rotation function Z-score (in PHASER)
RIP	radiation-induced phasing
PDB	protein data bank
PEG	poly ethylene glycol
poly-Ala	alanine polypeptide
r.m.s.(d.)	root mean square (deviation)
RNase	ribonuclease
(S-)SAD	(native sulfur-based) single wavelength anomalous diffraction
SIR(AS)	single isomorphous replacement (with anomalous scattering)
sof	site occupation factor
s.u.	standard uncertainty
TFZ	translation function Z-score (in PHASER)
v/v	volume per volume
w/v	weight per volume
vs.	*versus*
e.g.	*exemplo gratia*

Bibliography

Acquati, F., Morelli, C., Cinquetti, R., Bianchi, M. G., Porrini, D., Varesco, L., Gismondi, V., Rocchetti, R., Talevi, S., Possati, L., Magnanini, C., Tibiletti, M. G., Bernasconi, B., Daidone, M. G., Shridhar, V., Smith, D. I., Negrini, M., Barbanti-Brodano, G., & Taramelli, R. (2001). Cloning and characterization of a senescence inducing and class II tumor suppressor gene in ovarian carcinoma at chromosome region 6q27. Oncogene *20*(8), 980–988.

Acquati, F., Possati, L., Ferrante, L., Campomenosi, P., Talevi, S., Bardelli, S., Margiotta, C., Russo, A., Bortoletto, E., Rocchetti, R., Calza, R., Cinquetti, R., Monti, L., Salis, S., Barbanti-Brodano, G., & Taramelli, R. (2005). Tumor and metastasis suppression by the human RNASET2 gene. Int J Oncol *26*(5), 1159–1168.

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., & Zwart, P. H. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution.. Acta Crystallogr D Biol Crystallogr *66*(Pt 2), 213–221.

Afonine, P. V. (2010). On atomic displacement parameters (ADP) and their parameterization in PHENIX. Computational Crystallography Newsletter *1*, 24–31.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*(17), 3389–3402.

Arnold, K., Bordoli, L., Kopp, J., & Schwede, T. (2006). The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics *22*(2), 195–201.

Astbury, W. T. (1947). X-ray studies of nucleic acids.. Symp Soc Exp Biol (1), 66–76.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D., & Zardecki, C. (2002). The Protein Data Bank. Acta Crystallogr D Biol Crystallogr *58*(Pt 6 No 1), 899–907.

Blobel, G. & Dobberstein, B. (1975). Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. J Cell Biol *67*(3), 835–851.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., & Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res *31*(1), 365–370.

Brünger, A. T. (1992). Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature *355*(6359), 472–475.

Bruker (2003). SAINT, Bruker AXS Inc., Madison, Wisconsin, USA.

Campomenosi, P., Salis, S., Lindqvist, C., Mariani, D., Nordström, T., Acquati, F., & Taramelli, R. (2006). Characterization of RNASET2, the first human member of the Rh/T2/S family of glycoproteins. Arch Biochem Biophys *449*(1-2), 17–26.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., & Richardson, D. C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr *66*(Pt 1), 12–21.

Chen, Y. W., Dodson, E. J., & Kleywegt, G. J. (2000). Does NMR Mean Not for Molecular Replacement? Using NMR-Based Search Models to Solve Protein Crystal Structures. Structure *8*(11), R213 – R220.

Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. EMBO J *5*(4), 823–826.

Claude, J.-B., Suhre, K., Notredame, C., Claverie, J.-M., & Abergel, C. (2004). CaspR: a web server for automated molecular replacement using homology modelling. Nucleic Acids Res *32*(Web Server issue), W606–W609.

Collaborative Computational Project (1994). The CCP4 suite: programs for protein crystallography. Acta Crystallogr D Biol Crystallogr *50*(Pt 5), 760–763.

Collaborative Computational Project (2011). CCP4 wiki (ccp4wiki.org).

Consortium, U. (2011). Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res *39*(Database issue), D214–D219.

Debreczeni, J. E., Bunkóczi, G., Girmann, B., & Sheldrick, G. M. (2003a). In-house phase determination of the lima bean trypsin inhibitor: a low-resolution sulfur-SAD case. Acta Crystallogr D Biol Crystallogr *59*(Pt 2), 393–395.

Debreczeni, J. E., Bunkóczi, G., Ma, Q., Blaser, H., & Sheldrick, G. M. (2003b). In-house measurement of the sulfur anomalous signal and its use for phasing. Acta Crystallogr D Biol Crystallogr *59*(Pt 4), 688–696.

Deshpande, R. A. & Shankar, V. (2002). Ribonucleases from T2 family. Crit Rev Microbiol *28*(2), 79–122.

Didisheim, J.-J. & Schwarzenbach, D. (1987). Rigid-link constraints and rigid-body molecules. Acta Crystallographica Section A *43*(2), 226–232.

Dieckmann, S. (2009). Studien zur Pathogenese humaner Leukoenzephalopathien http://webdoc.sub.gwdg.de/diss/2009/diekmann/.

Diederichs, K. & Karplus, P. A. (1997). Improved R-factors for diffraction data analysis in macromolecular crystallography. Nat Struct Biol *4*(4), 269–275.

Einspahr, H. M. & Weiss, M. S. (2011). Quality Indicators in Macromolecular Crystallography: Definitions and Applications. submitted for International Tables of Crystallography Volume F.

Emsley, P. & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr *60*(Pt 12 Pt 1), 2126–2132.

Emsley, P., Lohkamp, B., Scott, W. G., & Cowtan, K. (2010). Features and development of Coot. Acta Crystallogr D Biol Crystallogr *66*(Pt 4), 486–501.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., & Bairoch, A. (2003). ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res *31*(13), 3784–3788.

Guex, N., Peitsch, M. C., & Schwede, T. (2009). Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective. Electrophoresis *30 Suppl 1*, S162–S173.

Gupta, R., Brunak, S., & Brunak, S. R. (2002). Prediction of Glycosylation Across the Human Proteome and the Correlation To Protein Function. Pacific Symposium on Biocomputing *7*, 310–322.

Harada, Y., Lifchitz, A., Berthou, J., & Jolles, P. (1981). A translation function combining packing and diffraction information: an application to lysozyme (high-temperature form). Acta Crystallographica Section A *37*(3), 398–406.

Hardman, K. D. & Ainsworth, C. F. (1972). Structure of concanavalin A at 2.4-Å resolution. Biochemistry *11*(26), 4910–4919.

Hendrickson, W. A., Smith, J. L., & Sheriff, S. (1985). Direct phase determination based on anomalous scattering. Methods Enzymol *115*, 41–55.

Henneke, M., Diekmann, S., Ohlenbusch, A., Kaiser, J., Engelbrecht, V., Kohlschütter, A., Krätzner, R., Madruga-Garrido, M., Mayer, M., Opitz, L., Rodriguez, D., Rüschendorf, F., Schumacher, J., Thiele, H., Thoms, S., Steinfeld, R., Nürnberg, P., & Gärtner, J. (2009). RNASET2-deficient cystic leukoencephalopathy resembles congenital cytomegalovirus brain infection. Nat Genet *41*(7), 773–775.

Holm, L. & Sander, C. (1997). Dali/FSSP classification of three-dimensional protein folds. Nucleic Acids Res *25*(1), 231–234.

Hooft, R. W., Vriend, G., Sander, C., & Abola, E. E. (1996). Errors in protein structures. Nature *381*(6580), 272.

Irie, M. (1999). Structure-function relationships of acid ribonucleases: lysosomal, vacuolar, and periplasmic enzymes. Pharmacol Ther *81*(2), 77–89.

Irie, M., Ohgi, K., Iwama, M., Koizumi, M., Sasayama, E., Harada, K., Yano, Y., Udagawa, J., & Kawasaki, M. (1997). Role of histidine 46 in the hydrolysis and the reverse transphosphorylation reaction of RNase Rh from Rhizopus niveus. J Biochem *121*(5), 849–853.

Johnson, C. K. (1965). OR TEP-II: A Fortran Thermal Ellipsoid Plot Program for Crystal Structure Illustrations QRNL-5138 Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Kabsch, W. (2010). XDS. Acta Crystallographica Section D *66*(2), 125–132.

Karle, J. (1980). Some Developments in Anomalous Dispersion for the Structural Investigation of Macromolecular Systems in Biology. International Journal of Quantum Chemistry: Quantum Biology Symposium *7*, 357–367.

Kiefer, F., Arnold, K., Künzli, M., Bordoli, L., & Schwede, T. (2009). The SWISS-MODEL Repository and associated resources. Nucleic Acids Res *37*(Database issue), D387–D392.

Kleywegt, G. J. (1996). Use of non-crystallographic symmetry in protein structure refinement. Acta Crystallogr D Biol Crystallogr *52*(Pt 4), 842–857.

Krissinel, E. & Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr *60*(Pt 12 Pt 1), 2256–2268.

Krissinel, E. & Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. J Mol Biol *372*(3), 774–797.

Liu, Y., Emilion, G., Mungall, A. J., Dunham, I., Beck, S., Meuth-Metzinger, V. G. L., Shelling, A. N., Charnock, F. M. L., & Ganesan, T. S. (2002). Physical and transcript map of the region between D6S264 and D6S149 on chromosome 6q27, the minimal region of allele loss in sporadic epithelial ovarian cancer. Oncogene *21*(3), 387–399.

Luhtala, N. & Parker, R. (2010). T2 Family ribonucleases: ancient enzymes with diverse roles. Trends Biochem Sci *35*(5), 253–259.

Martoglio, B. & Dobberstein, B. (1998). Signal sequences: more than just greasy peptides. Trends Cell Biol *8*(10), 410–415.

Massa, W. (2007). (Vieweg+Teubner Verlag). 5th edition edit.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., & Read, R. J. (2007). Phaser crystallographic software. J Appl Crystallogr *40*(Pt 4), 658–674.

Mesters, J. R. & Hilgenfeld, R. (2007). Protein Glycosylation, Sweet to Crystal Growth. Crystal Growth & Design *7*(11), 2251–2253.

Milbradt, A. G., Kerek, F., Moroder, L., & Renner, C. (2003). Structural characterization of hellethionins from helleborus purpurascens. Biochemistry *42*(8), 2404–2411.

Mitsui, Y. & Wyckoff, H. W. (1975). The crystal struture of monoclinic ribonuclease-S at six Ångstroms resolution. J Mol Biol *94*(1), 17–31.

Murshudov, G. (2010). personal communication.

Murshudov, G. N., Vagin, A. A., & Dodson, E. J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr *53*(Pt 3), 240–255.

Ohashi, Y., Mitsuhara, I., Oshshima, M., Ugaki, M., Hirochika, H., Honkura, R., Iwai, T., & Nakamura, S. (2001). Method for producing disease resistant plant with thionin gene from Avena sativa - United States Patent 6187995.

Oszlányi, G. & Süto, A. (2004). Ab initio structure solution by charge flipping.. Acta Crystallogr A *60*(Pt 2), 134–141.

Painter, J. & Merritt, E. A. (2006). Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. Acta Crystallogr D Biol Crystallogr *62*(Pt 4), 439–450.

Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S., & Tucker, P. A. (2009). On the combination of molecular replacement and single-wavelength anomalous diffraction phasing for automated structure determination. Acta Crystallogr D Biol Crystallogr *65*(Pt 10), 1089–1097.

Panjikar, S. & Tucker, P. A. (2002). Xenon derivatization of halide-soaked protein crystals. Acta Crystallogr D Biol Crystallogr *58*(Pt 9), 1413–1420.

Pape, T. & Schneider, T. R. (2004). HKL2MAP: a graphical user interface for macromolecular phasing with SHELX programs. Journal of Applied Crystallography *37*(5), 843–844.

Rabijns, A., Verboven, C., Rougé, P., Barre, A., Damme, E. J. M. V., Peumans, W. J., & Ranter, C. J. D. (2002). Structure of an RNase-related protein from Calystegia sepium. Acta Crystallogr D Biol Crystallogr *58*(Pt 4), 627–633.

Raines, R. T. (1998). Ribonuclease A. Chem Rev *98*(3), 1045–1066.

Ramachandran, G. N. & Sasisekharan, V. (1968). Conformation of polypeptides and proteins. Adv Protein Chem *23*, 283–438.

Read, R. J., Bunkoczi, G., McCoy, A., & Oeffner, R. (2006). *PHASER 2.0 manual*. Cambridge Institute for Medical Research Hills Road, Cambridge, CB2 0XY UK.

Rodríguez, D. D., Grosse, C., Himmel, S., González, C., de Ilarduya, I. M., Becker, S., Sheldrick, G. M., & Usón, I. (2009). Crystallographic ab initio protein structure solution below atomic resolution. Nat Methods *6*(9), 651–653.

Rodriguez, R., Chinea, G., Lopez, N., Pons, T., & Vriend, G. (1998). Homology modeling, model and software evaluation: three related resources. Bioinformatics *14*(6), 523–528.

Roeser, D., Dickmanns, A., Gasow, K., & Rudolph, M. G. (2005). De novo calcium/sulfur SAD phasing of the human formylglycine-generating enzyme using in-house data. Acta Crystallogr D Biol Crystallogr *61*(Pt 8), 1057–1066.

Rollett, J. S. (1970). *Crystallographic Computing* pp. 167–181 (F. R. Ahmed and S. R. Hall and C. P. Huber: Copenhagen: Munksgaard.).

Rosenfield, Jnr, R. E., Trueblood, K. N., & Dunitz, J. D. (1978). A test for rigid-body vibrations based on a generalization of Hirshfeld's 'rigid-bond' postulate. Acta Crystallographica Section A *34*(5), 828–829.

Roversi, P., Johnson, S., & Lea, S. M. (2010). With phases: how two wrongs can sometimes make a right. Acta Crystallogr D Biol Crystallogr *66*(Pt 4), 420–425.

Rupp, B. (2009). (Garland Science). 1 edit.

Schneider, T. R. (1996). What can we learn from anisotropic temperature factors. in *SERC Daresbury Laboratory, Daresbury* pp. 133–144.

Schneider, T. R. & Sheldrick, G. M. (2002). Substructure solution with SHELXD. Acta Crystallogr D Biol Crystallogr *58*(Pt 10 Pt 2), 1772–1779.

Schomaker, V. & Trueblood, K. N. (1968). On the rigid-body motion of molecules in crystals. Acta Crystallographica Section B *24*(1), 63–76.

Schuermann, J. P. & Tanner, J. J. (2003). MRSAD: using anomalous dispersion from S atoms collected at Cu Kalpha wavelength in molecular-replacement structure determination. Acta Crystallogr D Biol Crystallogr *59*(Pt 10), 1731–1736.

Sheldrick, G. M. (1990). Phase annealing in SHELX-90: direct methods for larger structures. Acta Crystallographica Section A *46*(6), 467–473.

Sheldrick, G. M. (2002). Macromolecular phasing with SHELXE. Zeitschrift fuer Kristallographie *217*(12-2002), 644–650.

Sheldrick, G. M. (2008). A short history of SHELX. Acta Crystallogr A *64*(Pt 1), 112–122.

Sheldrick, G. M. (2009). SADABS - Bruker AXS area detector scaling and absorption correction Version 2009/1.

Sheldrick, G. M. (2010). Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. Acta Crystallogr D Biol Crystallogr *66*(Pt 4), 479–485.

Sheldrick, G. M. (2011). XPREP - Bruker data preparation and recoiprocal space exploration Version 2011/1.

Smirnoff, P., Roiz, L., Angelkovitch, B., Schwartz, B., & Shoseyov, O. (2006). A recombinant human RNASET2 glycoprotein with antitumorigenic and antiangiogenic characteristics: expression, purification, and characterization. Cancer *107*(12), 2760–2769.

Stein, N. (2008). *CHAINSAW*: a program for mutating pdb files used as templates in molecular replacement. Journal of Applied Crystallography *41*(3), 641–643.

Stuart, D. I. & Phillips, D. C. (1985). On the derivation of dynamic information from diffraction data. Methods Enzymol *115*, 117–142.

Sumner, J. (1926). The isolation and crystallization of the enzyme urease. Preliminary paper. Journal of Biological Chemistry *69*, 435–41.

Tanaka, N., Arai, J., Inokuchi, N., Koyama, T., Ohgi, K., Irie, M., & Nakamura, K. T. (2000). Crystal structure of a plant ribonuclease, RNase LE. J Mol Biol *298*(5), 859–873.

Tickle, I. J. (2007). Experimental determination of optimal root-mean-square deviations of macromolecular bond lengths and angles from their restrained ideal values. Acta Crystallographica Section D *63*(12), 1274–1281.

Tickle, I. J., Laskowski, R. A., & Moss, D. S. (2000). R(free) and the R(free) ratio. II. Calculation Of the expected values and variances of cross-validation statistics in macromolecular least-squares refinement. Acta Crystallogr D Biol Crystallogr *56*(Pt 4), 442–450.

Trueblood, K. N., Bürgi, H.-B., Burzlaff, H., Dunitz, J. D., Gramaccioli, C. M., Schulz, H. H., Shmueli, U., & Abrahams, S. C. (1996). Atomic Dispacement Parameter Nomenclature. Report of a Subcommittee on Atomic Displacement Parameter Nomenclature. Acta Crystallographica Section A *52*(5), 770–781.

Urzhumtseva, L., Afonine, P. V., Adams, P. D., & Urzhumtsev, A. (2009). Crystallographic model quality at a glance. Acta Crystallogr D Biol Crystallogr *65*(Pt 3), 297–300.

Usón, I. & Sheldrick, G. M. (1999). Advances in direct methods for protein crystallography. Curr Opin Struct Biol *9*(5), 643–648.

van Aalten, D. M., Bywater, R., Findlay, J. B., Hendlich, M., Hooft, R. W., & Vriend, G. (1996). PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules.. J Comput Aided Mol Des *10*(3), 255–262.

Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. J Mol Graph *8*(1), 52–6, 29.

Weiss, M. S. (2001). Global indicators of X-ray data quality. Journal of Applied Crystallography *34*(2), 130–135.

Weiss, M. S. & Hilgenfeld, R. (1997). On the use of the merging *R* factor as a quality indicator for X-ray data. Journal of Applied Crystallography *30*(2), 203–205.

Winn, M. D., Isupov, M. N., & Murshudov, G. N. (2001). Use of TLS parameters to model anisotropic displacements in macromolecular refinement. Acta Crystallogr D Biol Crystallogr *57*(Pt 1), 122–133.

Yoshida, H. (2001). The ribonuclease T1 family. Methods Enzymol *341*, 28–41.

Acknowledgements

I would like to express my deep gratitude to Prof. George M. Sheldrick, who was the best 'Doktorvater' I could have hoped for. He provided everything needed by me and for this thesis to grow. Most of the work described here was based on his programs and algorithms.

In all but the last two months of this work, I was accompanied everyday by Dr. Christian Große, who contributed in so many ways. I also owe thanks to the other members of the Sheldrick lab:

Dr. Regine Herbst-Irmer for her generous help at all times.

Dr. Tim Grüne for his critical and extremely helpful remarks. Together with Tim and Carlos Eduardo da Cunha, Hellethionin D was crystallized and measured.

Dr. Tobias Beck and Navdeep Sidhu for many valuable discussions.

The kind Dalila Griffin and our technician Helmut Dehnhard, who introduced me to diffractometer maintenance.

I have to thank Dr. Birger Dittrich who acts as second advisor for this thesis. He always challenged me for good explanations and had a friendly word for me when needed most. It was a pleasure to work shoulder to shoulder to him and his group – Dr. Julian Holstein, Dr. Christian B. Hübschle and of course Kevin Pröpper.

Many people from other research facilities contributed as well:

Dr. Ralph Krätzner and Dr. Dr. Robert Steinfeld kindly provided me with the purified human RNase T2 and many ideas regarding the crystallization of this protein. Their co-worker Mark Ziegenbein assisted me in the crystallization. Dr. Stefan Becker made the crystallization robots available to us and gave important tips on both NMR and crystallization.

Prof. Dr. Isabel Usón has modified and run ARCIMBOLDO for the Hellethionin D structure; in summer 2009, she also introduced me to PHASER and SHELXE.

Prof. Dr. Henning Urlaub, Uwe Plessmann and He-Hsuan Hsiao performed the mass spectrometry on human RNase T2.

I dedicate this work to my family:

To my fiancé Florian, who is always by my side – especially in the final days of this work. I may return the favour soon.

I am deeply indebted to our parents, giving us endless encouragement and support for the pursue of science – and of life. Especially my father, Dr. Volker Thorn, in whose footsteps I walk.

And to my three beloved brothers Christian, Raed and Izak.

Finally, I am grateful for my friends – they accompanied me all along – Christine Rauscher, Anne Rubbert, Jost Menzel, Stella Duck and Oliver König.

Curriculum Vitae

Name	**Andrea Regina Shirin Thorn**
Birth	21.12.1982 in Hamburg, Germany
Parents	Nega Thorn and Dr. Volker Thorn
Education	
2002	**Abitur**, Labenwolf Gymnasium Nuremberg, Germany
2002 – 2005	**B. Sc. (Molecular Science)**, Friedrich-Alexander University Erlangen Thesis: 'Structure and Synthesis of Transition Metal Complexes with tripodal N,N,O Ligands', supervisor: Prof. Dr. Nicolai Burzlaff
2005	**Research stay**, University of Chiba, Japan Molecular biology study of vascular smooth muscle cells
2005 – 2007	**M. Sc. (Molecular Life Science)**, Friedrich-Alexander University Erlangen Thesis: 'Structure and function of 5β-Progersterone Reductase from Digitalis lanata', supervisor: Prof. Dr. Yves Muller
2008 – 2011	**Doctoral dissertation**, Georg-August-University Goettingen Thesis: 'Practical approaches to macromolecular X-ray structure determination', supervisor: Prof. George M. Sheldrick FRS
Work	
2002	**Temporary lab assistant**, Heumann Pharma (Pfizer), Feucht Quality Control for pharmaceutical production
from 2008	**Teaching assistant**, Georg-August University Goettingen Conception and supervision of practical courses for chemistry students and graduate students in the Molecular Biology and Neurosciences (GGNB) faculty; exercise tutoring.
Other	**Voluntary assistant** for the Lindau-Nobel foundation (2006-2009)