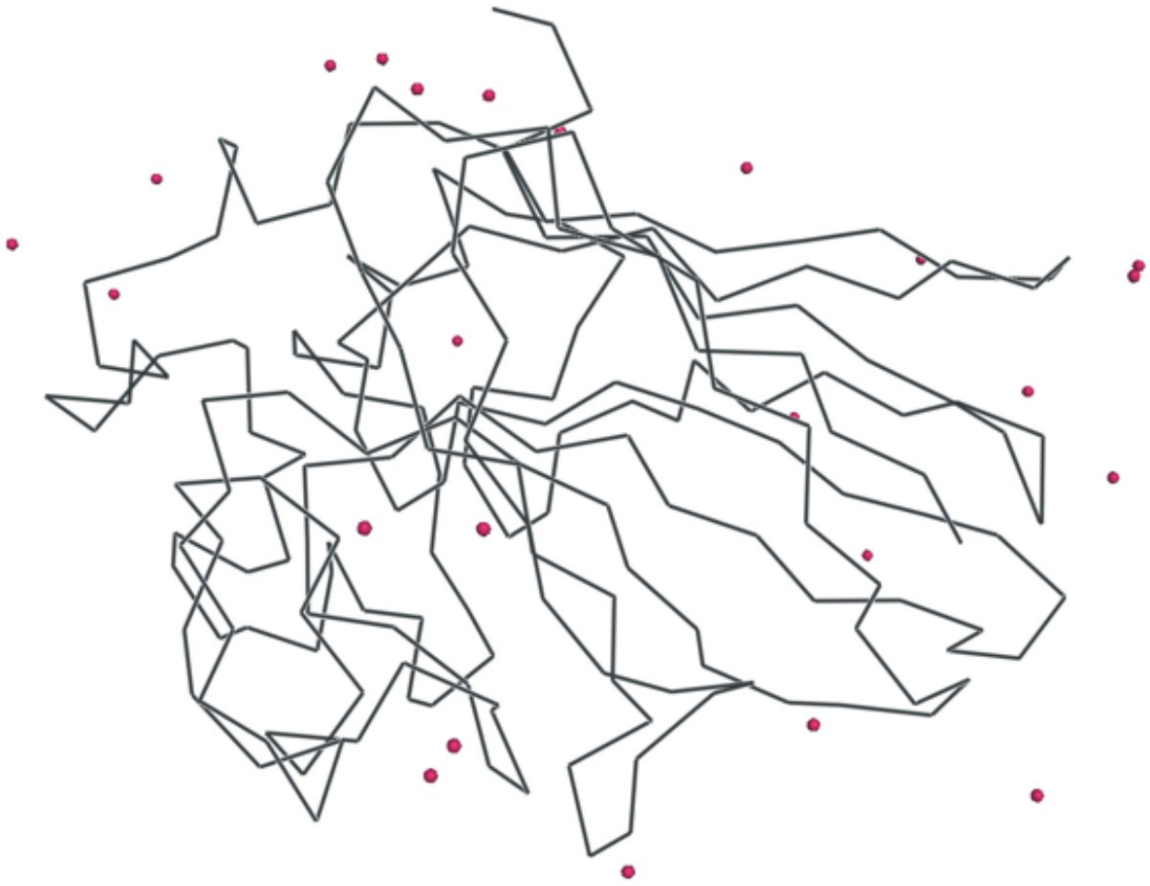
A tutorial for learning and teaching macromolecular crystallography

Annette Faust, Santosh Panjikar, Uwe Mueller, Venkataraman Parthasarathy, Andrea Schmidt, Victor S. Lamzin and Manfred S. Weiss



Reference: Faust *et al*. (2008). J. Appl. Cryst. (in press).

Experiment 2: MAD on bromide-soaked Thaumatin

Thaumatin is a mixture of three intensely sweet proteins isolated from the seed vessel of the Katemfe plant (*Thaumatococcus daniellii*). It is about 1,000 times sweeter than sucrose on a weight basis and 100,000 times on a molar basis and it is therefore used in food industry as a sweetener. The commercially available thaumatin is a mixture of thaumatin I and thaumatin II (Figure 1) with traces of other sweet proteins. The amino acid sequence of thaumatin contains 207 residues, where thaumatin I and II differ from one another in five amino acids only. As it is a mixture, it is hard to examine the ratio between thaumatin I and II in the crystal structure during refinement. All PDB-entries of thaumatin (Figure 2b) are therefore modelled using the thaumatin I sequence (Ko *et al.*, 1994).

Over the past two decades, multiple wavelength anomalous diffraction (MAD) has been the standard method for a *de novo* structure determination in macromolecular crystallography (Hendrickson and Ogata; 1997; Hendrickson, 1999). In MAD, the wavelength dependent anomalous scattering properties from heavy atoms that are part of or bound to the macromolecule of interest are utilized. The heavy atoms can directly be incorporated in the protein (e.g. seleno-methionine derivatives or metal-containing proteins) or they can be soaked into the crystal. MAD experiments are carried out at different X-ray energies around an absorption edge of the heavy atom where the anomalous scattering factors of the heavy atom are significantly different from each other. Up to four diffraction data sets are collected at the peak wavelength where Δf'' reaches its maximum, at the inflection point wavelength where Δf' reaches its minimum and away from the absorption edge at wavelengths of the high energy or/and low energy side, at least 100 eV remote from the peak.

Sequence alignment of Thaumatin I and II:

```
I     ATFEIVNR CSYTVWAAAS KGDAALDAGG RQLNSGESWT INVEPGTNGG KIWARTDCYF
II    ATFEIVNR CSYTVWAAAS KGDAALDAGG RQLNSGESWT INVEPGTKGG KIWARTDCYF

I   DDSGSGICKT GDCGGLLRCK RFGRPPTTLA EFSLNQYGKD YIDISNIKGF NVPMNFSPTT
II  DDSGRGICRT GDCGGLLQCK RFGRPPTTLA EFSLNQYGKD YIDISNIKGF NVPMDFSPTT

I   RGCRGVRCAA DIVGQCPAKL KAPGGGCNDA CTVFQTSEYC CTTGKCGPTE YSRFFKRLCP
II  RGCRGVRCAA DIVGQCPAKL KAPGGGCNDA CTVFQTSEYC CTTGKCGPTE YSRFFKRLCP

I   DAFSYVLDKP TTVTCPGSSN YRVTFCPTA
II  DAFSYVLDKP TTVTCPGSSN YRVTFCPTA
```

Figure 1: Amino acid sequence alignment of thaumatin I and II.

1 Crystallisation

Chemicals: Thaumatin (M≈ 22.2 KDa/mol, Sigma-Aldrich cat.no. T-7638)
N-(2-acetamido)iminodiacetic acid (ADA) (M= 190.15 g/mol, Sigma cat.no.A9883)
L(+)-Tartraric acid potassium sodium salt (M = 282.22 g/mol, Sigma-Aldrich cat.no. S6170)
Glycerol (M =92.09 g/mol, Sigma-Aldrich cat.no. G9012)
Milli-Q water

Thaumatin crystals grew within a few days after mixing 2 µl of of protein solution (15 mg/ml in 0.1 M ADA pH 6.5) and 2 µl of reservoir solution (0.1 M ADA pH 6.5, 1.0 M sodium/potassium tartrate) and equilibrating the drop against the reservoir. The tetragonal crystals (space group $P4_12_12$, space group number 92) exhibit unit cell parameters a =57.7 Å, c= 150.2 Å and diffract X-rays to beyond 1.5 Å resolution (Müller-Dieckmann *et al*., 2005). Two representative crystallization experiments are shown in Figure 2. Prior to data collection, the crystals were derivatized by soaking them for a few seconds in a solution containing 1 M sodium bromide and 25% (v/v) glycerol and flash-cooled in liquid nitrogen.

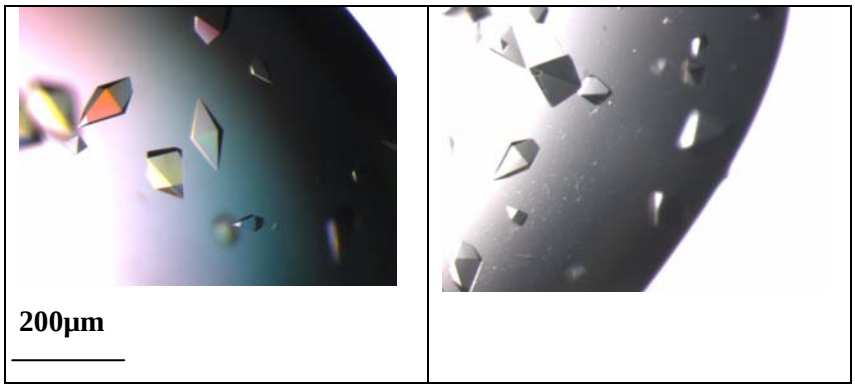


Figure 2: Tetragonal Thaumatin crystals.

2 Data Collection

Prior to diffraction data collection, an X-ray fluorescence scan at the bromine K-absorption edge was performed. For this the X-ray fluorescence of the sample is measured as a function of beam energy (Figure 3). The anomalous signal as expressed by the anomalous scattering length Δf'' goes along the extent of X-ray fluorescence. The dispersive signal can be derived from the anomalous signal by means of the Kramers-Kronig equation. The wavelengths for data collection were derived from this scan. This can either be done manually or using the program CHOOCH (Evans and Pettifer, 2001). The wavelengths chosen were those at the peak of the absorption edge where Δf'' reaches its maximum value, at the inflection point where Δf' reaches its minimum values and at least 100 eV away from the absorption edge at wavelengths of the high energy and low energy side of the absorption edge. Based on the fluorescence scan, the peak wavelength was chosen as λ = 0.91878 Å (E = 13,494 eV), the inflection point wavelength as λ = 0.91960 Å (E = 13,482 eV), and the high and low energy remote wavelengths as λ = 0.91337 Å and 0.92523 Å (E = 13,574 and 13,400 eV), respectively. X-ray diffraction data have then been collected at the tunable beam line X12 at the EMBL Hamburg Outstation (DESY Hamburg). The beamline is equipped with a MARMosaic-CCD detector (225mm) from MARRESEARCH (Norderstedt, Germany) and a MARdtb goniostat (MARRESEARCH, Norderstedt, Germany).

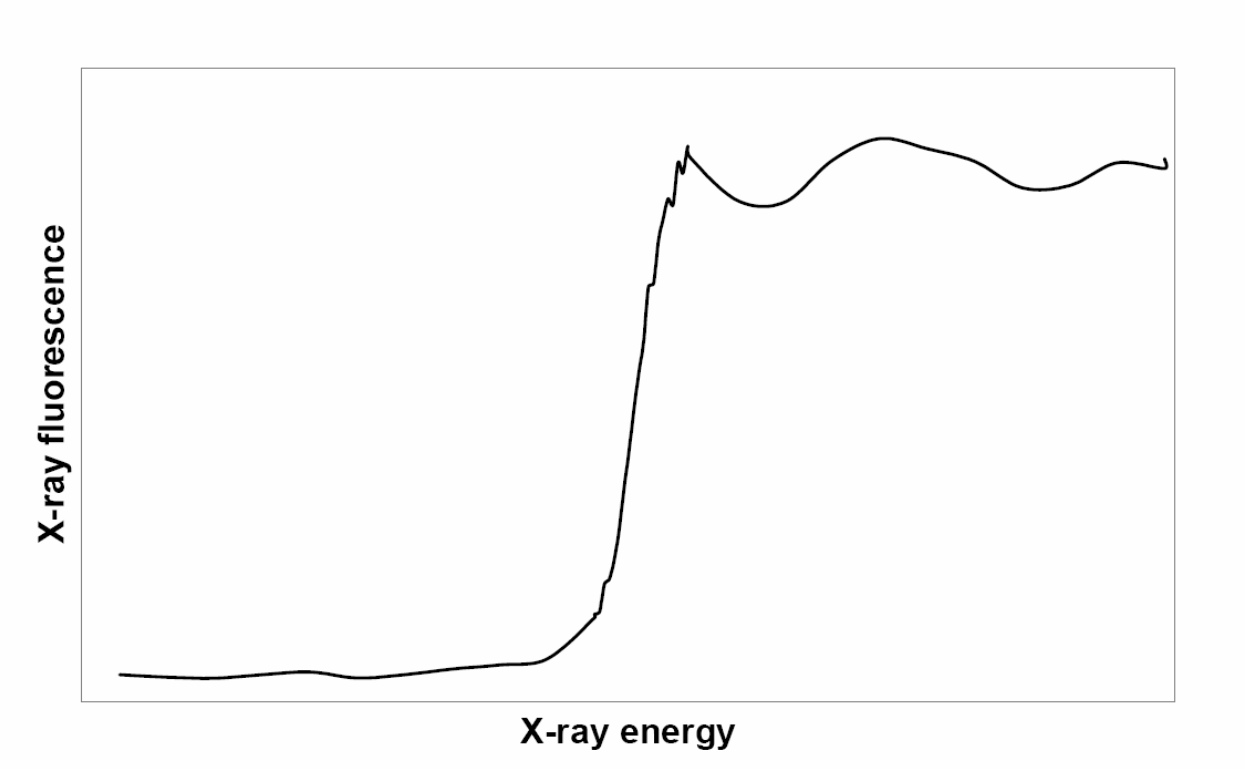


Figure 3: X-ray fluorescence scan around the K-absorption edge of Br. The energy range of the scan is approximately from 13430 to 13550 eV.

The relevant data collection parameters are given below:

All data sets:	detector distance:	200 mm
	oscillation range/image:	0.5º
	no. of images:	360 for all 4 data sets
peak:	path to images:	experiment2/data/peak/
	wavelength	0.91878 Å
	exposure time/frame:	4.5 sec
	image names:	thau_peak_1_###.mccd
inflection:	path to images:	experiment2/data/infl/
	wavelength	0.91960 Å
	exposure time/frame:	4.1 sec
	image names:	thau_inf_1_###.mccd
high energy remote:	path to images:	experiment2/data/hrem/
	wavelength	0.91337 Å
	exposure time/frame:	4.3 sec
	image name:	thau_hirem_1_###.mccd
low energy remote:	path to images:	experiment2/data/lrem/
	wavelength:	0.92523 Å
	exposure time/frame:	4.3 sec
	images name:	thau_lorem_1_###.mccd

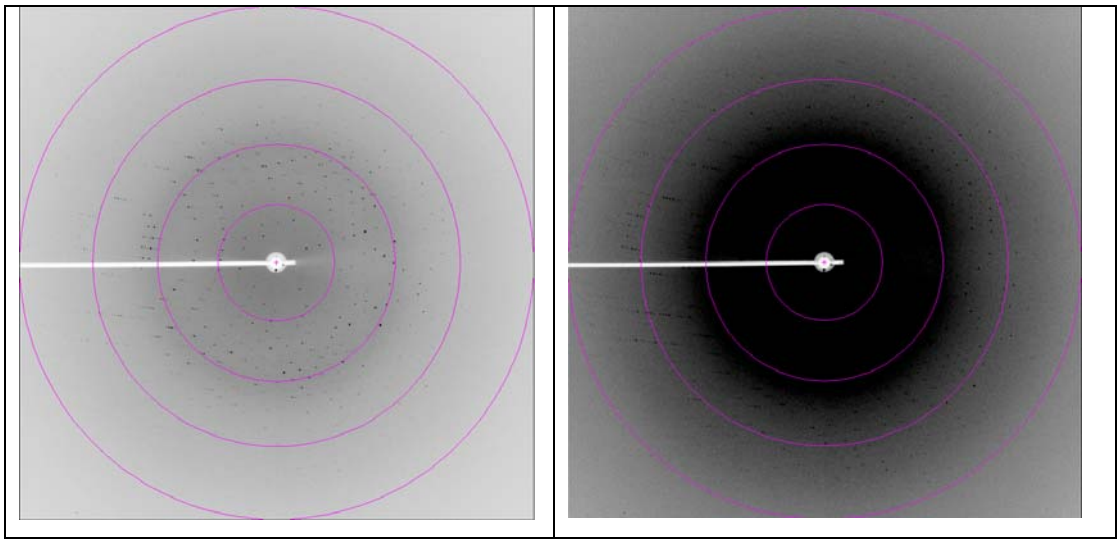


Figure 4: Diffraction image of tetragonal thaumatin at the peak wavelength displayed using different contrast levels. The resolution rings are shown at 7.3, 3.6, 2.4 and 1.8 Å, respectively.

3 Data Processing

The data were indexed, integrated and scaled using the program XDS (Kabsch, 1993). XDS is able to recognize compressed images, therefore it is not necessary to unzip the data before using XDS. (For use with other programs this will be necessary and can be done using the command **bunzip2** *.bz2). XDS needs only one input file. This has to be called XDS.INP, no other name is recognised by the program. In XDS.INP the image name given must not include the zipping-format extension (*.img instead of *.img.bz2). Further, XDS has a very limited string length (80) to describe the path to the images. Therefore it may be necessary to create a soft link to the directory containing the images by using the command **ln -s** /path/to/images/ ./images. The path to the images in XDS.INP will then be ./images/.

- **indexing** **1st run of XDS**

Before running XDS, the XDS.INP file has to be edited so that it contains the correct data collection parameters. To estimate the space group and the cell parameters the space group number in XDS.INP has to be set to 0. These parameters will be obtained in the output file IDXREF.LP.

JOBS= XYCORR INIT COLSPOT IDXREF space group number=0

XYCORR computes a table of spatial correction values for each pixel

INIT determines an initial background for each detector pixel and finds the trusted region of the detector surface.

COLSPOT collects strong diffraction spots from a specified subset of the data images

IDXREF interprets observed spots by a crystal lattice and refines all diffraction parameters.

The IDXREF.LP output file contains the results of the indexing. For thaumatin, the correct space group is $P4_12_12$ (space group number 92) with unit cell parameters a=57.7 c=150.2Å.

- **integration** **2nd run of XDS**

After determination of space group and cell parameters all images will be integrated and corrections (radiation damage, absorption, detector etc.) will be calculated in a second XDS run.

DEFPIX defines the trusted region of the detector, recognizes and removes shaded areas, and eliminates regions outside the resolution range defined by the user.

XPLAN helps planning data collection. Typically, one or a few data images are collected initially and processed by XDS. XPLAN reports the completeness of data that could be expected for various starting angles and total crystal rotation.

Warning: If data were initially processed for a crystal with unknown cell constants and space group, the reported results will refer to space group P1.

INTEGRATE collects 3-dimensional profiles of all reflections occurring in the data images and estimates their intensities.

CORRECT corrects intensities for decay, absorption and variations of detector surface sensitivity, reports statistics of the collected data set and refines the diffraction parameters using all observed spots.

The file CORRECT.LP contains the statistics for the complete data set after integration and corrections. After truncation a file name XDS_ASCII.HKL will be written out, which contains the integrated and scaled reflections. If the cell parameters and the space group are known already one can run XDS with JOBS=ALL.

- **scaling run XSCALE**

The collected images have to be on a common scale. The correction factors are determined and applied to compensate absorption effects and radiation damage. Individual reflections can be corrected for radiation damage (0-dose corrections). XSCALE writes out a *.ahkl file, which can be converted with XDSCONV to be used within the CCP4-suite (Collaborative Computational Project, 1994) or other programs.

Table 1: Data processing statistics (from XSCALE.LP)

	peak	**inflection**	**high energy remote**	**low energy remote**
Resolution limits [Å]	10.0 - 1.82			
Unit cell parameters				
a, b, c [Å]	57.8, 57.8, 150.0			
Space group	$P4_12_12$			
Mosaicity [°]	0.15	0.24	0.24	0.24
Total number of reflections	330728	330340	335120	324832
Unique reflections	43279	43238	43467	42524
Completeness [%]	99.5 (98.2)	99.4 (97.9)	99.9 (100.0)	97.8 (91.2)
Redundancy	7.6 (7.3)	7.6 (7.3)	7.7 (7.6)	7.6 (7.3)
I/σ(I)	14.0 (4.2)	15.3 (4.6)	14.7 (4.3)	18.3 (5.9)
$R_{r.i.m.}$ / R_{meas} [%]	13.1 (50.4)	11.6 (45.5)	12.6 (50.6)	9.0 (34.9)
Wilson B-factor	19.2	19.8	19.8	20.8

- **converting *.ahkl to *.mtz run XDSCONV with XDSCONV.INP**

XDSCONV.INP:

```
OUTPUT_FILE=thau_peak.mtz CCP4
INPUT_FILE=thau_peak.ahkl
```

XDSCONV creates an input file F2MTZ.INP for the final conversion to binary mtz-format. To run the CCP4 programs F2MTZ and CAD, just type the two commands:

```
f2mtz HKLOUT temp.mtz < F2MTZ.INP
cad HKLIN1 temp.mtz HKLOUT thau_peak_ccp4.mtz << EOF
LABIN FILE 1 ALL
END
EOF
```

Alternatively the XDS_ASCII.HKL file can be converted with COMBAT (to mtz-format) and this mtz-file can be used as an input file for SCALA in CCP4.

4 Structure Solution

The structure was solved using the 3W-MAD protocol of AUTO-RICKSHAW: the EMBL-Hamburg automated crystal structure determination platform (Panjikar *et al*., 2005). AUTO-RICKSHAW can be accessed from outside EMBL under www.embl-hamburg.de/AutoRickshaw/LICENSE (a free registration may be required, please follow the instructions on the web page). AUTO-RICKSHAW uses the file XDS_ASCII.HKL directly. In the following the automatically generated summary of AUTO-RICKSHAW is printed together with the results of the structure determination:

The input diffraction data were prepared and converted for use in AUTO-RICKSHAW using programs of the CCP4 suite (Collaborative Computational Project, 1994). F_A values were calculated using the program SHELXC (Sheldrick *et al*., 2001). Based on an initial analysis of the data, the maximum resolution for substructure determination and initial phase calculation was set to 2.4 Å. All of the 20 heavy atoms requested were found using the program SHELXD (Schneider and Sheldrick, 2002). The correct hand for the substructure was determined using the programs ABS (Hao, 2004) and SHELXE (Sheldrick, 2002). Initial phases were calculated after density modification using the program SHELXE (Sheldrick, 2002). 98.54% of the model was built using the program ARP/wARP (Perrakis *et al.*, 1999; Morris *et al.*, 2004). The model was further refined using COOT (Emsley and Cowtan, 2004) and REFMAC5 (Murshudov *et al.*, 1997). First, the missing parts of the model were built and a bound tartrate molecule was identified and included. An anomalous difference Fourier map was calculated to identify all of the bromide ions. Figures 5-7 illustrate the progress in structure solution. Figure 5 shows the experimental map after density modification followed by initial model building in ARP/wARP 7.0. Figure 6 shows the refined structure with the corresponding ($2F_{obs}$-F_{calc},α_{calc})-electron density map superimposed and Figure 7 with the anomalous difference Fourier electron density map superimposed. The peaks in Figure 7 identify the bromide ions bound to the structure.

More details can be found in the AUTORICKSHAW output (directory: experiment2/autorickshaw). The AUTORICKSHAW run in the advanced version using three wavelength data (peak, inflection, high energy remote) took 47 minutes.

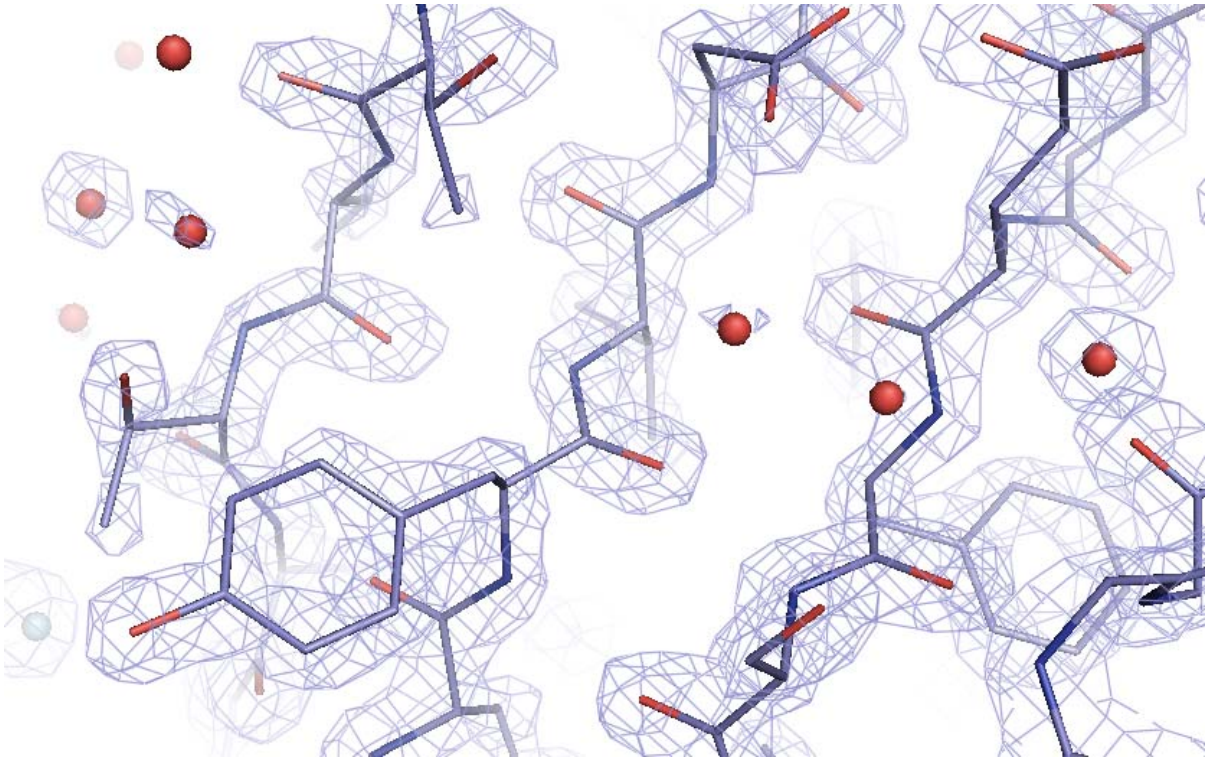


Figure 5: Initial experimental electron density map after density modification and initial model building contoured at 2.0σ.

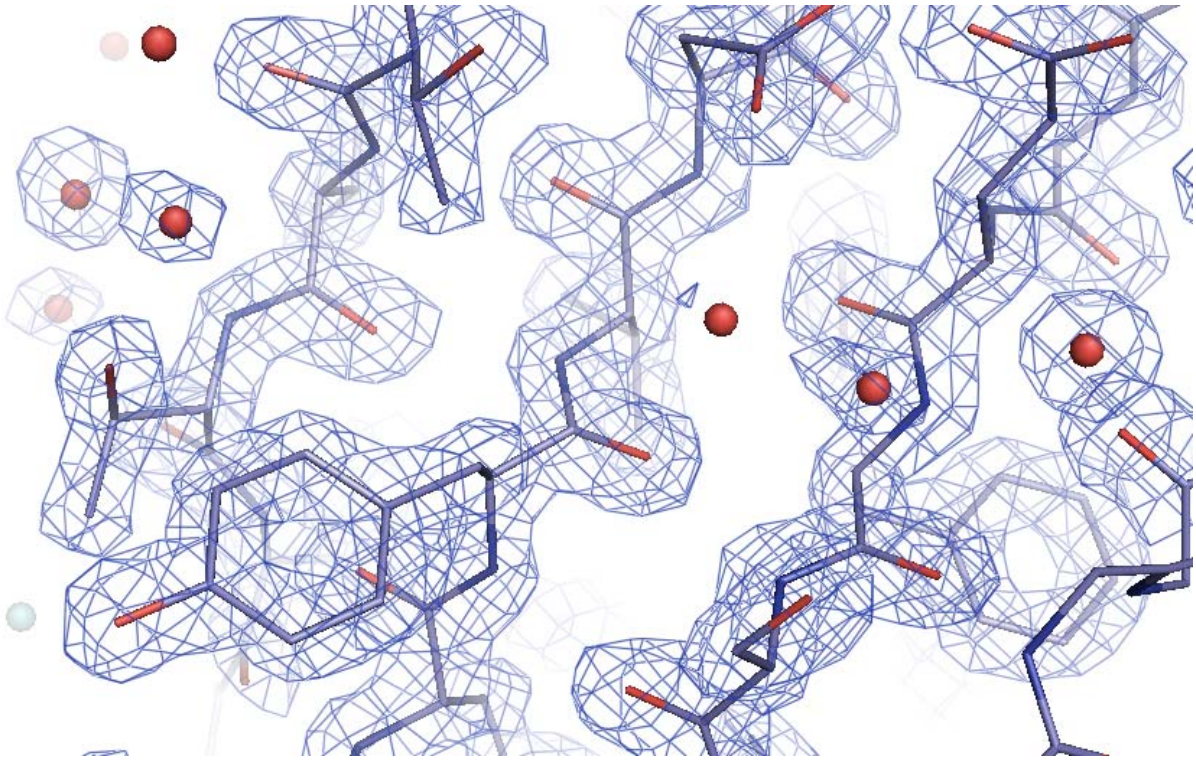


Figure 6: Refined structure and the corresponding ($2F_{obs}$-F_{calc},α_{calc})-electron density map contoured at 2.0σ.

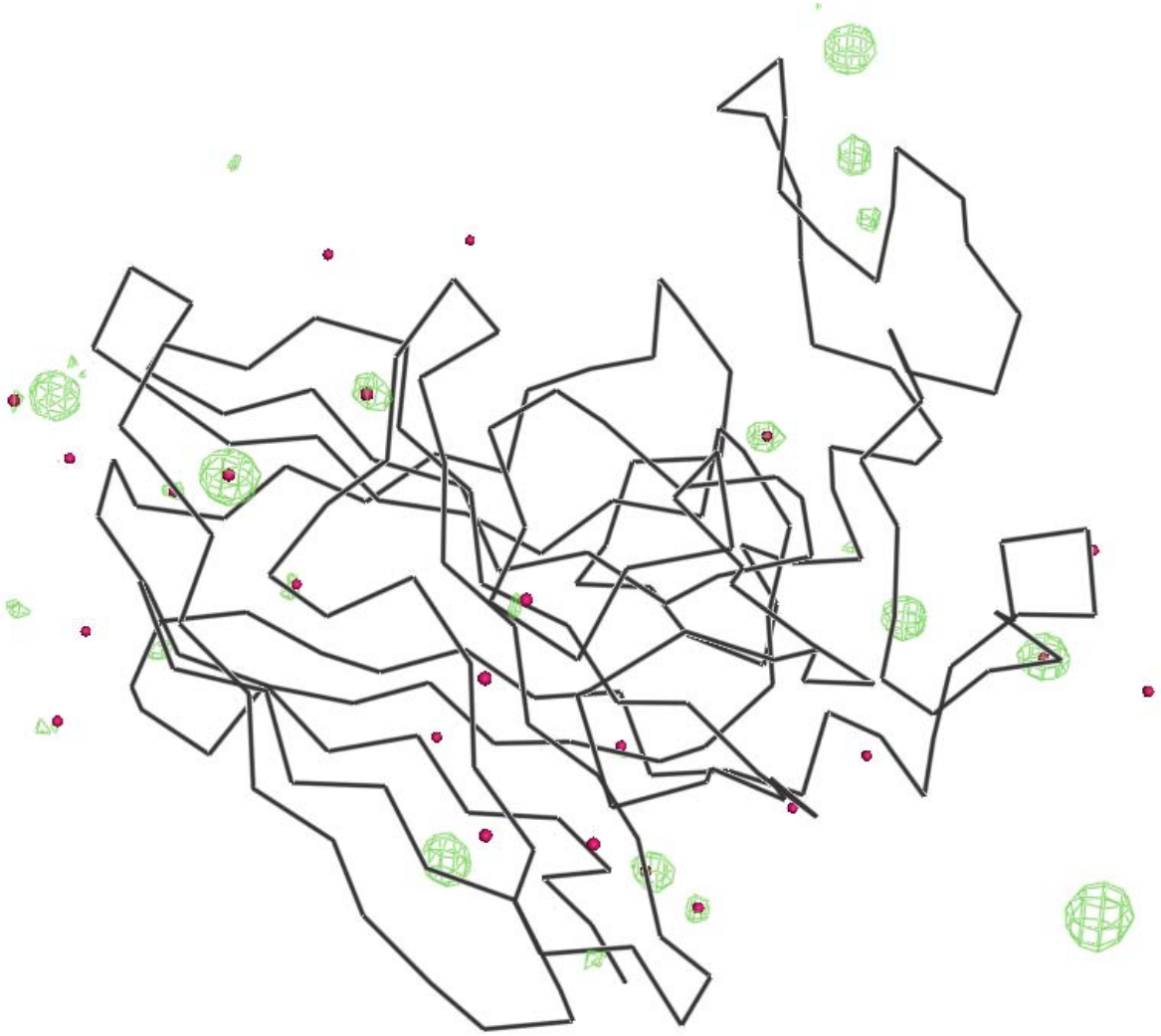


Figure 7: Anomalous difference Fourier electron density map superimposed to a C_α-representation of the final model of thaumatin. The bromide ions observed to be bound to the surface of thaumatin are shown as red spheres. The map is contoured at 4.0 σ.

5 References

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* **D50**, 760-763.

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

Emsley P., Cowtan, K. (2004). *Acta Cryst.* **D60**, 2126-2132.

Evans, G., & Pettifer, R.F. (2001). *J. Appl. Cryst.* **34**, 82-86.

Hao, Q. (2004). *J. Appl. Cryst.* **37**, 498-499.

Hendrickson, W.A., and Ogata, C.M. (1997). *Methods Enzymol.* **276**, 494-522.

Hendrickson, W.A. (1999). *J. Synchr. Rad.* **6**, 845-851.

Kabsch, W. (1993). *J. Appl. Cryst.* **26**, 795-800.

Ko, T.-P., Day, J., Greenwood, A. & McPherson, A. (1994). *Acta Cryst.* **D50**, 813-825.

La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472-494.

Mueller-Dieckmann, C., Panjikar, S., Tucker, P. A. & Weiss, M. S. (2005). *Acta Cryst*. **D61**, 1263-1272.

Morris, R. J., Perrakis, A. & Lamzin, V. S. (2002). *Acta Cryst*. **D58**, 968-975.

Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S. & Tucker, P. A. (2005). *Acta Cryst.* **D61**, 449-457.

Perrakis, A., Morris, R. J. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458-463.

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

Sheldrick, G. M., Hauptman, H. A., Weeks, C. M., Miller, R. & Uson, I. (2001). *International Tables for Macromolecular Crystallography,* **Vol. F**, edited by M. G. Rossmann & E. Arnold, ch. 16, pp. 333-345. Dordrecht: Kluwer Academic Publishers.

Sheldrick, G. M. (2002). *Z. Kristallogr.* **217**, 644-650.