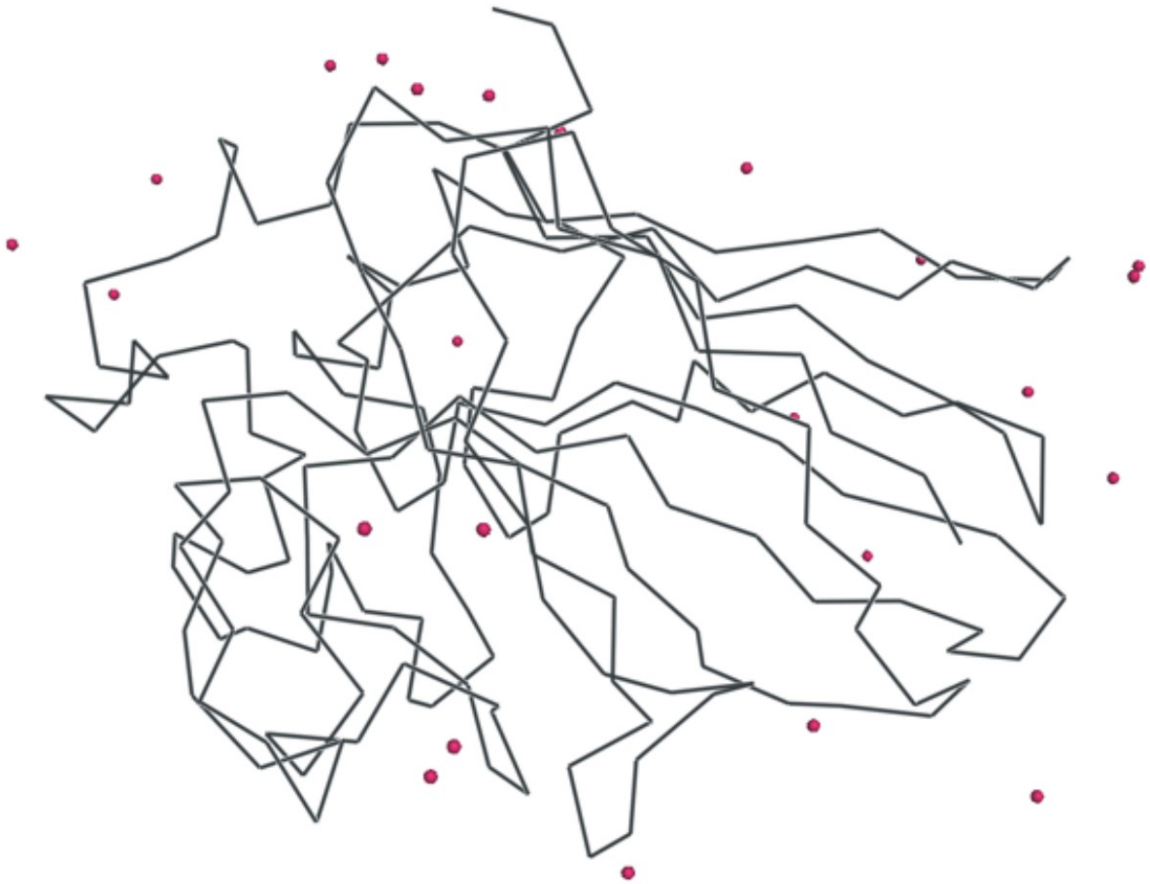
A tutorial for learning and teaching macromolecular crystallography – version 2010

Annette Faust, Sandra Puehringer, Nora Darowski, Santosh Panjikar, Venkataraman Parthasarathy, Andrea Schmidt, Victor S. Lamzin, Kay Diederichs, Uwe Mueller and Manfred S. Weiss



References: A. Faust *et al*. (2008). *J. Appl. Cryst.* **41**, 1161-1172.

A. Faust *et al*. (2010). *J. Appl. Cryst.* **43** (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 are therefore modeled using the thaumatin I sequence (Ko *et al.*, 1994). Thaumatin also contains eight disulfide bonds.

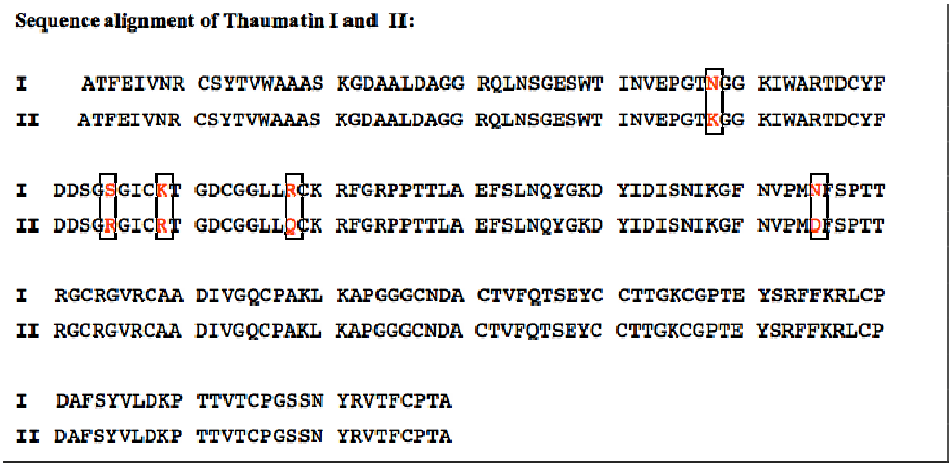


Figure 1: Amino acid sequence alignment of thaumatin I and II.

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 the anomalous scattering length $\Delta f''$ reaches its maximum, at the inflection point wavelength where the dispersive scattering length $\Delta f'$ 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.

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.

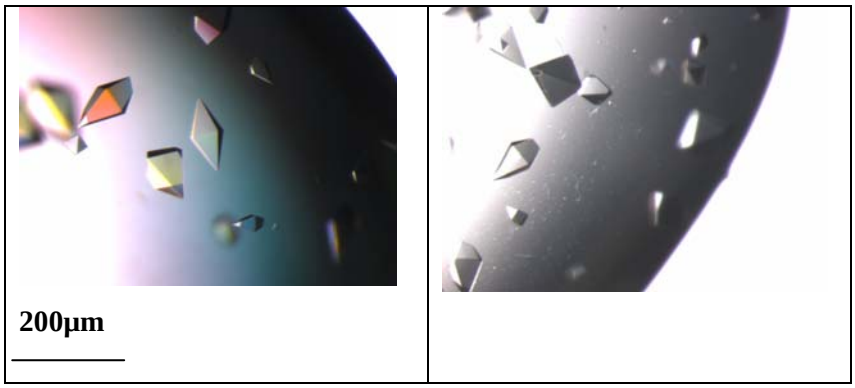


Figure 2: Tetragonal Thaumatin crystals.

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 incident beam energy (Figure 3). The anomalous signal as expressed by the anomalous scattering length Δf'' goes along with the magnitude of the 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 & 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).

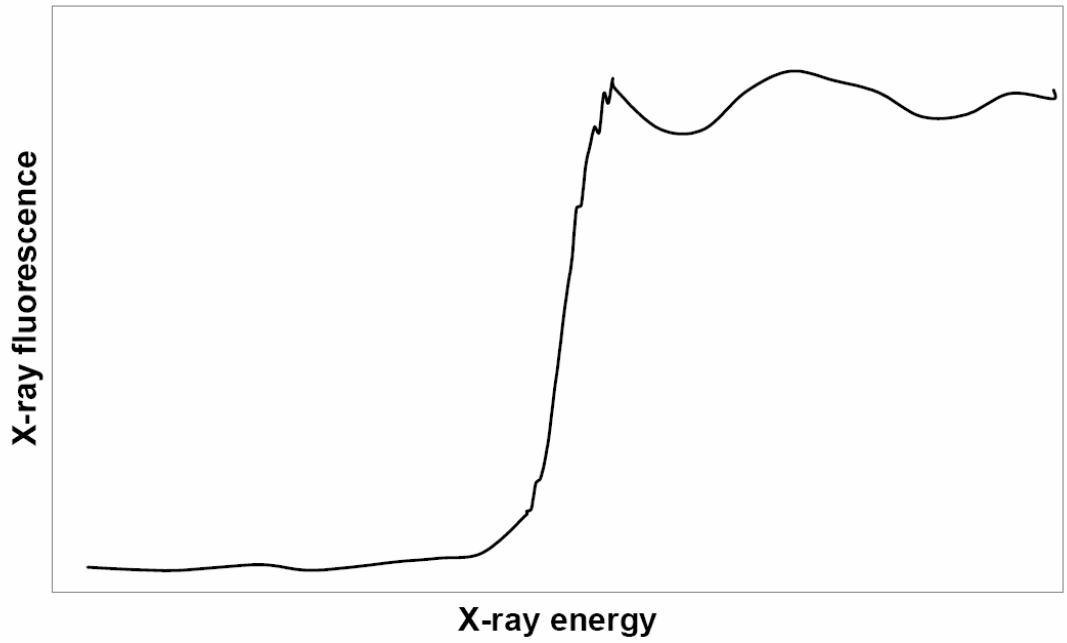


Figure 3: X-ray fluorescence scan around the K-absorption edge of Br. The energy range of the scan is approximately from 13430 to 13550 eV.

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:	exp2/data/peak/
	wavelength	0.91878 Å
	exposure time/frame:	4.5 sec
	image names:	exp2_thau_mad_peak_1_###.img
inflection:	path to images:	exp2/data/infl/
	wavelength	0.91960 Å
	exposure time/frame:	4.1 sec
	image names:	exp2_thau_mad_infl_1_###.img
high energy remote:	path to images:	exp2/data/hrem/
	wavelength	0.91337 Å
	exposure time/frame:	4.3 sec
	image name:	exp2_thau_mad_hrem_1_###.img
low energy remote:	path to images:	exp2/data/lrem/
	wavelength:	0.92523 Å
	exposure time/frame:	4.3 sec
	images name:	exp2_thau_mad_lrem_1_###.img

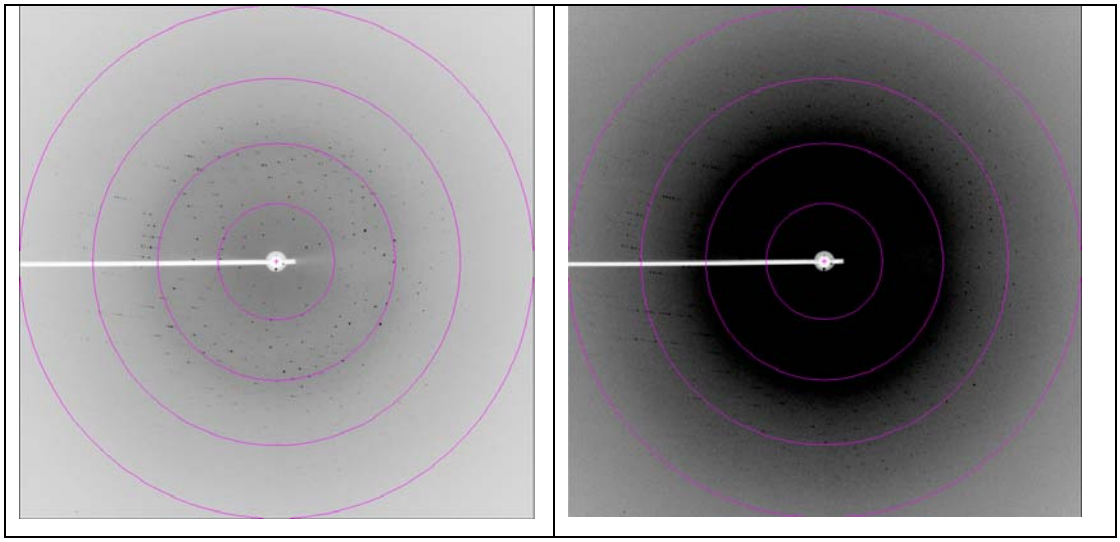


Figure 4: Diffraction image of tetragonal thaumatin at the peak wavelength displayed using different contrast levels. The resolution rings are shown at 7.3, 3.6, 2.4 and 1.8 Å, respectively.

3 Data Processing

The collected diffraction data were indexed, integrated and scaled using the program XDS (Kabsch, 1993, 2010a,b). XDS is simply run by the command ***xds***. If a multi-processor machine is available, the command ***xds_par*** can be used, which calls a parallel version of XDS and consequently runs much faster. XDS needs only one input file, which must be called XDS.INP. No other name is recognized by the program. The file XDS.INP contains all relevant information about the data collection, from beam parameters to detector parameters and crystal parameters (if known) as well as the data collection geometry. In XDS.INP one can also define the steps through which the program should go. This is done by using the parameter JOBS. The following command, which is equivalent to JOBS= ALL would make XDS run through all eight steps XYCORR, INIT, COLSPOT, IDXREF, DEFPIX, XPLAN, INTEGRATE and CORRECT.

```
JOBS= XYCORR INIT COLSPOT IDXREF DEFPIX XPLAN INTEGRATE CORRECT
```

In the XYCORR step, tables of spatial correction factors are set up (if required). INIT calculates the gain of the detector and produces an initial background table. COLSPOT identifies strong reflections which are used for indexing. IDXREF performs the actual indexing of the crystal. DEFPIX identifies the regions on the detector surface which are used for measuring intensities, XPLAN helps to devise a data collection strategy, INTEGRATE integrates the reflection intensities of the whole data set and CORRECT scales and merges symmetry-related reflections and multiple measurements. It also prints out data processing statistics. After completing each individual step, a log-file with a name corresponding to the step (*STEP-name*.LP) is written.

Action 1: edit the supplied file XDS.INP and insert the relevant information about the data collection, namely the data collection wavelength, crystal-to-detector distance, the direct beam coordinates, the total number of images, and rotation increment per image and of course very importantly the path to and the names of the image files. XDS is able to recognize compressed images; therefore it is not necessary to unzip the data before using XDS. 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/. If the space group and cell dimensions are known, the relevant information should be written into XDS.INP, if they are not known just set the parameter **SPACE GROUP NUMBER= 0**.

Action 2: run XDS until the indexing step, with the parameter JOBS set to:

```
JOBS= XYCORR INIT COLSPOT IDXREF
```

The output file IDXREF.LP contains the results of the indexing. It needs to be checked carefully whether the indexing is correct, since all subsequent steps assume the correctness of the indexing step. The most relevant parameters to look for are the STANDARD DEVIATION OF SPOT POSITION and the STANDARD DEVIATION OF SPINDLE POSITION. The first one should be in the order of 1 pixel, whereas the second one depends to some extent on the rotation increment per image but also on the mosaicity of the crystal. If it is 0.1º it is very good, if it is 0.5º it might still be ok, if it is larger than 1.0º the indexing has probably not worked. The table with the entries SUBTREE and POPULATION is also very interesting to look at. The first SUBTREE should have by a large margin more entries than all others. Also, the input parameters, such as the crystal-to-detector distance should after refinement not deviate too much from the input values.

The most common problem with the IDXREF step is that it often finishes with the message **!!! ERROR !!! INSUFFICIENT PERCENTAGE (<70%) OF INDEXED REFLECTIONS**. This means that less than 70% of the reflections that were collected in the COLSPOT step are not indexed, which may happen because of ice rings on the frames, split reflections or simply wrong input parameters. However, if all indicators of correct indexing are fine (see above) and no obvious errors can be identified then this message can be safely ignored and data processing can be continued.

If IDXREF was run with **SPACE_GROUP_NUMBER= 0**, an assumption of the correct Bravais lattice may be made at this stage. As a rule of thumb, choose the lattice of the highest possible symmetry, with a QUALITY OF FIT-value as low as possible (usually < 10). These numbers are printed in IDXREF.LP in the paragraph DETERMINATION OF LATTICE CHARACTER AND BRAVAIS LATTICE. Then, re-run the IDXREF step with the parameter SPACE_GROUP_NUMBER corresponding to the assumed Bravais lattice in XDS.INP. Alternatively, it is possible to make no assumption of the Bravais lattice at this point, and to simply continue with data integration. In this case the program will assume that space group determination should not be based on the metric symmetry of the lattice, but should be postponed to the CORRECT step (see below).

Nevertheless, if the true lattice is known, it should yield a good, i.e. low QUALITY OF FIT-number. For thaumatin, the correct space group is $P4_12_12$ (space group number 92) with unit cell parameters a = 57.7 c = 150.2Å.

Action 3: After the determination of the Bravais lattice and the cell parameters all images need to be integrated and corrections (radiation damage, absorption, detector etc.) will have to be calculated. This can be done in a further XDS run.

JOBS= DEFPIX XPLAN INTEGRATE CORRECT

The CORRECT step produces a file called CORRECT.LP, which contains the statistics for the complete data set after integration and corrections. For the statistics to be meaningful, the correct Laue symmetry has to be established first. To this end, the CORRECT step compares the statistics in all possible Laue groups. The correct Laue group is the one with the highest symmetry, which at the same time still exhibits an acceptable $R_{r.i.m.}/R_{meas}$. CORRECT writes a file named XDS_ASCII.HKL, which contains the integrated and scaled reflections.

The CORRECT step also performs a refinement of all geometric parameters and the cell dimensions based on all reflections of the data set. These parameters may be more accurate than the ones obtained from the indexing step. Therefore, one may try to use the refined parameters and to re-run the last XDS job. In order to not overwrite the original results, it is advisable to save all current files to a temporary directory. Then, the file GXPARM.XDS should be renamed or copied to XPARM.XDS and XDS be re-run. In case the original results are better, they can be copied back to the original directory.

While XDS will usually identify the correct Laue group, it does not determine the actual space group of the crystal. The decision about the existence of screw axes is left to the user. Indications, which screw axes may be present can be obtained from the table REFLECTIONS OF TYPE H,0,0 0,K,0 0,0,L OR EXPECTED TO BE ABSENT (*) in the file CORRECT.LP. Alternatively, the program POINTLESS (Evans, 2005) offers an automatic way of assigning the space group. POINTLESS can be run with the command ***pointless XDSIN XDS_ASCII.HKL.*** In the output the possible space groups together with their probabilities are given. Some space group ambiguity still remains at this stage, since it is impossible to distinguish between enantiomorphic space groups e.g. $P3_1$ and $P3_2$, or $P4_12_12$ and $P4_32_12$ just based on intensities. This ambiguity has to be resolved later during structure solution. The parameter SPACE_GROUP_NUMBER corresponding to the determined space group as well as the cell parameters should be entered into the file XDS.INP for running the next step.

Action 4: Finally, outlier reflections are identified by CORRECT by comparing their intensity to the average intensity in their respective resolution shells. These outliers may be removed, if there is a clear indication and reason for their existence: for example, ice rings often produce very strong reflections at specific d-spacings. The outliers are flagged as ‘alien’ in the file CORRECT.LP and their removal can simply be achieved by writing the outliers into a file called REMOVE.HKL. By re-running XDS with the command

JOBS= CORRECT

in XDS.INP, these outliers are then disregarded. This last action can be repeated until no more additional outliers are identified. However, the outlier removal has to be handled very carefully because strong reflections may also arise from non-crystallographic symmetry and in particular from the presence of pseudo-translation. A command to identify only the most extreme outliers would be ***awk '/alien/ { if (strtonum($5) > 19) print $0 }' CORRECT.LP >> REMOVE.HKL***. This command will remove outliers only when their Z-score is above 19. Hints to suitable criteria for outlier rejection can be found in the XDSwiki (http://strucbio.biologie.uni-konstanz.de/xdswiki), where this question is treated specifically in the article "Optimization".

Action 5:, The CORRECT step can be followed up by running the scaling program XSCALE, which is part of the XDS program package. This serves three purposes: a) the user may specify the limits of the resolution shells for which statistics should be printed, b) several XDS_ASCII.HKL files may be scaled together and c) correction factors for radiation damage may be applied to the data (see also the article "XSCALE" in the XDSwiki). XSCALE is run by simply typing ***xscale*** (or ***xscale_par*** to speed up the computation on a multi-processor machine) provided that a file XSCALE. INP defining the input and output files is present. As above in the CORRECT step, outliers may be rejected. 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.

Both CORRECT and XSCALE will produce all necessary output for assembling a table with all relevant data processing statistics, which is necessary for a publication.

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

Action 6: finally, the processed intensity file needs to be converted to certain file formats, which are used by other programs to perform the necessary structure determination steps. This can be achieved using the program XDSCONV, which can simply be run by using the command ***xdsconv*** provided that a file called XDSCONV.INP is present. XDSCONV.INP just needs to contain information about the name of the input file and about the name and type of the output file. If a CCP4-type file is required, XDSCONV reformats the reflection output file from XSCALE and creates an input file F2MTZ.INP for the final conversion of the reflection file to binary mtz-format, which is the standard format for all CCP4 programs (CCP4, 1994).

```
OUTPUT_FILE=thau_peak.hkl CCP4

INPUT_FILE=thau_peak.ahkl
```

To run the CCP4 programs F2MTZ and CAD, just type the two commands

```
f2mtz HKLOUT temp.mtz < F2MTZ.INP
```

and

```
cad HKLIN1 temp.mtz HKLOUT thau_peak_ccp4.mtz << eof
LABIN FILE 1 ALL
END
eof
```

Some CCP4 programs need the intensities of the Bijvoet pairs as input. For those the second parameter on the OUTPUT_FILE= line should be **CCP4_I** instead of **CCP4**. Alternatively, the file XDS_ASCII.HKL can be converted to mtz-format using the CCP4-programs COMBAT or POINTLESS (Evans, 2005) and this mtz-file can be used as an input file for the scaling program SCALA (Evans, 2005) in CCP4. More information on this can be found in the articles “Pointless” and “Scaling with SCALA” in the XDSwiki.

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, 2009). AUTO-RICKSHAW can be accessed from outside EMBL under www.embl-hamburg.de/Auto-Rickshaw/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 (directories: exp2/struct_sol and exp2/struct_ref). The AUTORICKSHAW run in the advanced version using three wavelength data (peak, inflection, high energy remote) took 47 minutes.

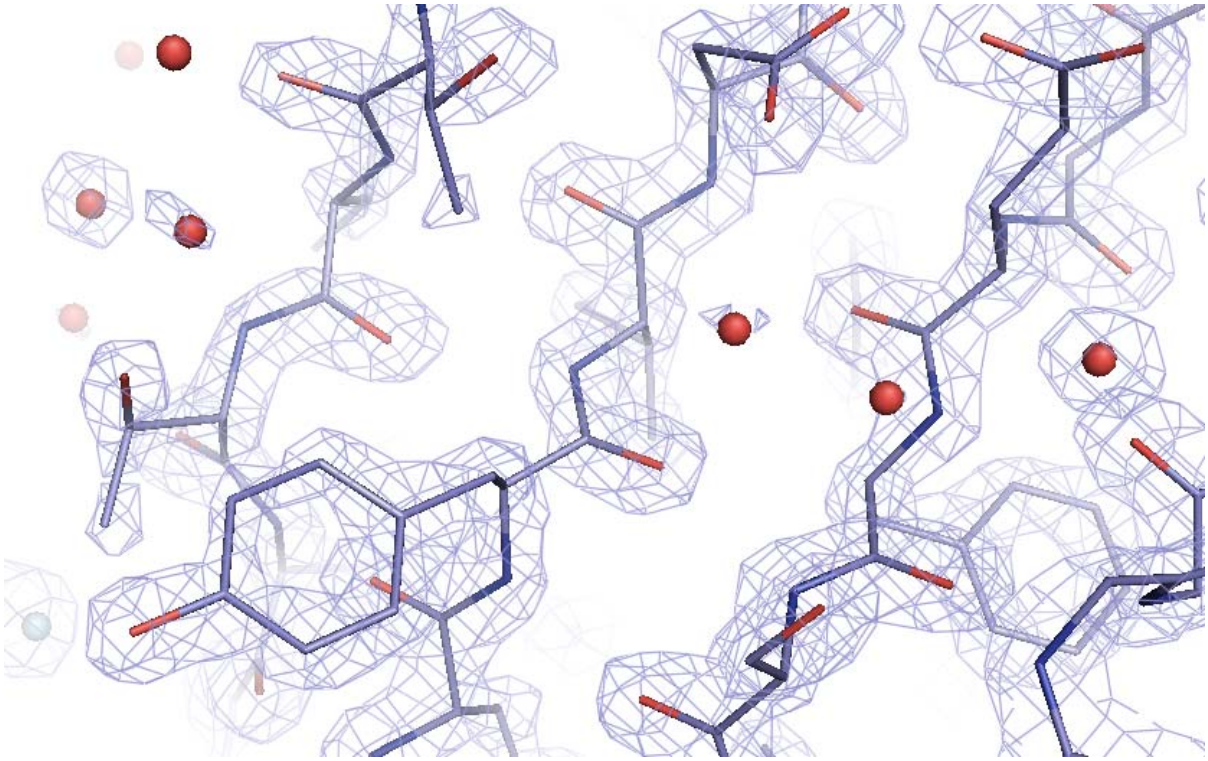


Figure 5: Initial experimental electron density map after density modification and initial model building contoured at 2.0σ.

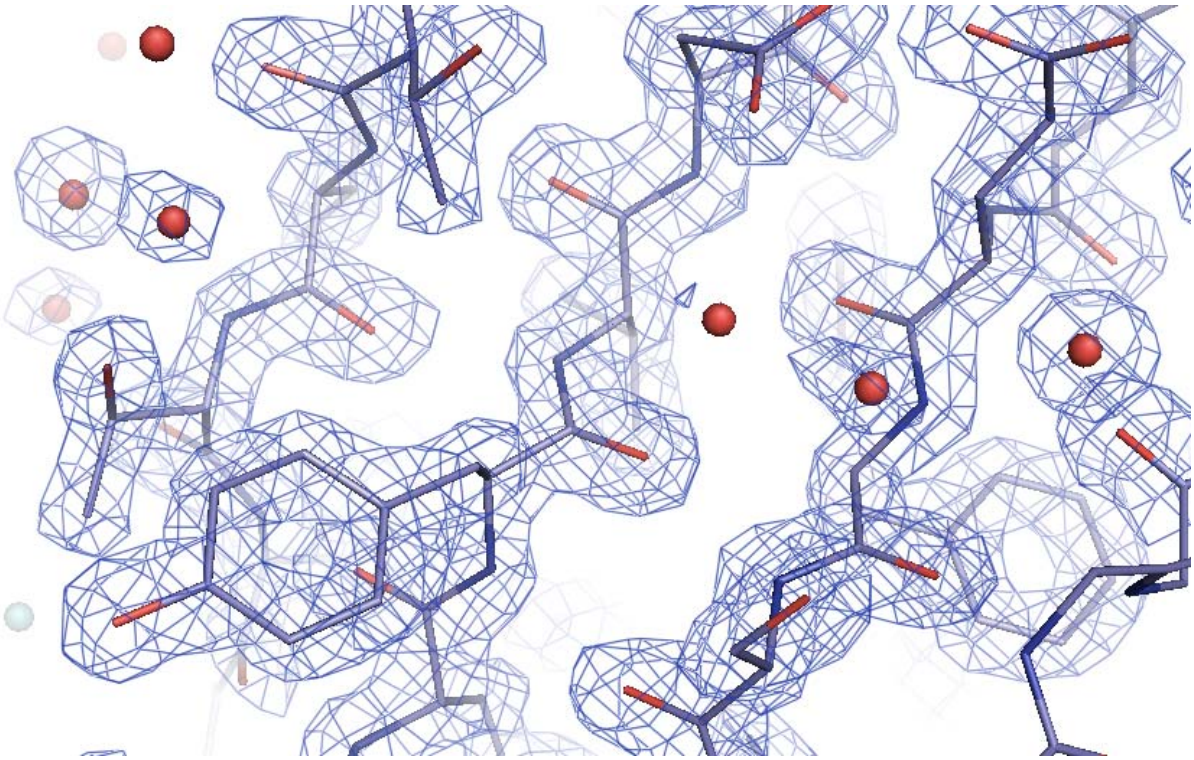


Figure 6: Refined structure and the corresponding ($2F_{obs}$-F_{calc}, α_{calc})-electron density map contoured at 2.0σ.

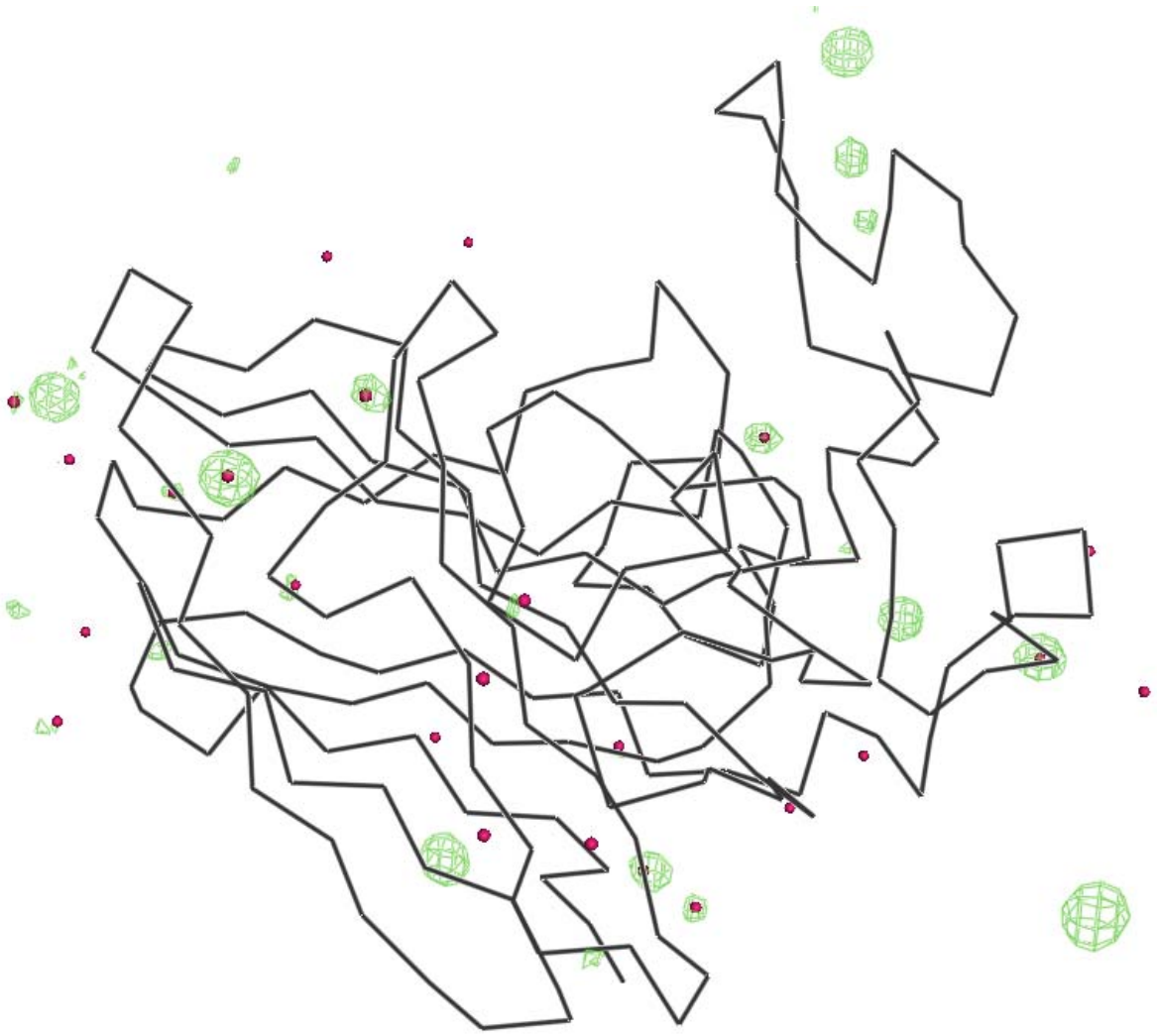


Figure 7: Anomalous difference Fourier electron density map superimposed onto a C_α-representation of the final model of thaumatin. The bromide ions observed to be bound to the surface of thaumatin are shown as red spheres. The map is contoured at 4.0 σ.

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.

Evans, P. (2005). Acta Cryst. **D62**, 72-82.

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

Hendrickson, W. A. & 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.

Kabsch, W. (2010a). *Acta Cryst.* **D66**, 125-132.

Kabsch, W. (2010b). *Acta Cryst.* **D66**, 133-144.

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.

Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S. & Tucker, P. A. (2009). *Acta Cryst.* **D65**, 1089-1097.

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.

Sheldrick, G. M. (2008). *Acta Cryst*. **A64**, 112-122.