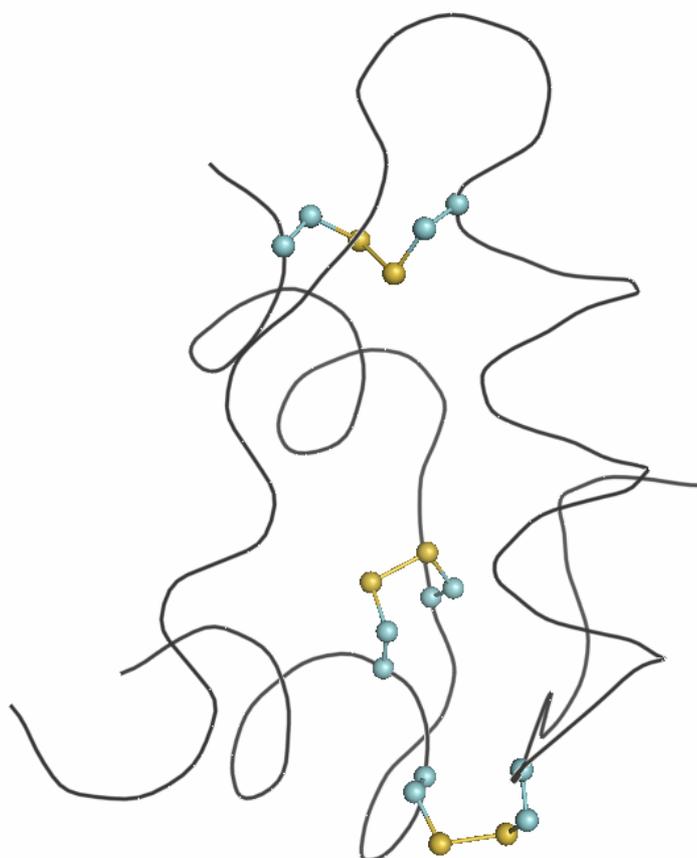# A tutorial for learning and teaching macromolecular crystallography – version 2010

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 1: S-SAD on bovine Insulin

Insulin regulates the cellular uptake, utilization, and storage of glucose, amino acids, and fatty acids and inhibits the breakdown of glycogen, protein, and fat. It is a two-chain polypeptide hormone produced by the β-cells of pancreatic islets (Voet *et al.*, 2006). The two chains comprise a total of 51 amino acids (M_W = 5,800 Da). The amino acid sequence is given in Figure 1. Three disulfide bonds hold the two chains together, one intra-chain SS-bridge between Cys6 and Cys11 in chain A and two inter-chain SS-bridges between Cys7 from chain A and Cys7 from chain B and Cys20 from chain A and Cys19 from chain B.

Experimental phase determination using single-wavelength anomalous diffraction (SAD) from the sulphur atoms inherently present in nearly all protein molecules has in the past few years experienced a huge boost in popularity. After the initial success with the small protein crambin (46 amino acids, 6 S-atoms) by Hendrickson and Teeter (1981), it took 18 years until the method was rediscovered by Dauter and colleagues (Dauter *et al.*, 1999), who were able to demonstrate that the structure of hen egg-white lysozyme (HEWL) could be successfully determined from the anomalous scattering of the protein S-atoms and surface-bound chloride ions alone. Since then, approximately 100 structures have been obtained using this so-called sulphur-SAD or S-SAD approach. Since the anomalous signal from the light atoms at the typically used wavelengths in macromolecular crystallography is small, it has been suggested and experimentally verified that the diffraction data collection at somewhat longer wavelengths may be beneficial (Djinovic-Carugo *et al.*, 2005; Mueller-Dieckmann *et al.*, 2005; Weiss *et al.*, 2001). However, while a larger anomalous signal may be obtained at longer X-ray wavelengths, additional experimental complications arising mainly from X-ray absorption have to be dealt with. In this experiment, cubic crystals of bovine insulin are used for experimental S-SAD phase determination using diffraction data collected at a wavelength of λ = 1.77 Å.
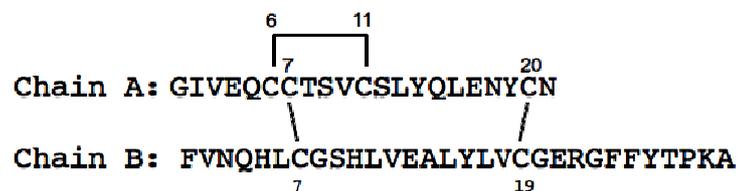

Figure 1: Amino acid sequence of bovine insulin including the three disulfide-bonds

1 Crystallisation

Chemicals: bovine insulin   (M = 5733.49 g/mol, Sigma cat. no. I5500)

Na$_2$HPO$_4$*12H$_2$O   (M = 358.14 g/mol, Fluka cat. no. 71649)

Na$_3$PO$_4$*12H$_2$O   (M = 380.12 g/mol, Fluka cat. no. 71908)

Na$_4$EDTA*4H$_2$O   (M = 452.23 g/mol, Fluka cat. no. 03699)

Na$_2$EDTA*2H$_2$O   (M = 372.24 g/mol Fluka cat. no. 03679)

Glycerol   (M = 92.09 g/mol, Sigma-Aldrich cat. no. G9012)

Milli-Q water

EasyXtal Tool screw-cap crystallization plates (Nextal, now Qiagen)

Bovine insulin crystals were prepared by the hanging drop-method in EasyXtal Tool crystallization plates. Unfortunately, these plates are not commercially available any longer. As a slightly less convenient replacement, greased Linbro plates with glass cover slides may be used. 4 μl of protein solution (20 mg/ml of protein dissolved in 20 mM Na$_2$HPO$_4$ and 10 mM Na$_3$EDTA pH 10.0–10.6) and 4 μl of reservoir solution (225-350 mM Na$_2$HPO$_4$/Na$_3$PO$_4$ pH 10.0-10.6, 10 mM Na$_3$EDTA) were mixed and equilibrated over reservoir solution. Crystals belonging to the cubic space group I2$_1$3 (space group number 199) with the unit-cell parameter a =7 8.0 Å grew within a few days up to a final size of 100-300 μm^3. They were cryo-protected in a solution containing 250 mM Na$_2$HPO$_4$/Na$_3$PO$_4$ pH 10.2, 10 mM Na$_3$EDTA, and 30% (v/v) glycerol, and usually diffracted X-rays to better than 1.4 Å.

200μm   200μm
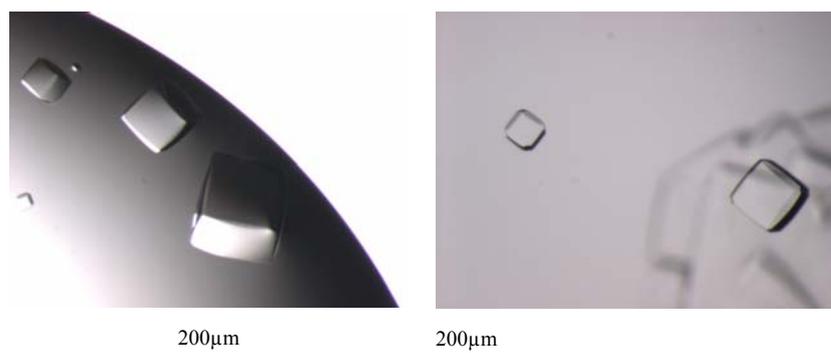
Figure 2: Cubic insulin crystals

2 Data collection

Diffraction data were collected at a wavelength of 1.77 Å at the tunable beam line BL 14.1 at the BESSY-II synchrotron in Berlin Adlershof. BL 14.1 is equipped with a MARMosaic-CCD detector (225mm) from the company MARRESEARCH (Norderstedt, Germany) and at the time of this experiment was equipped with a MARdtb goniostat (MARRESEARCH, Norderstedt, Germany).

The relevant data collection parameters are given below:

wavelength:	1.77 Å
detector distance:	50 mm
2θ-angle:	0°
oscillation range/image:	1.0°
no. of images:	180
path to images:	exp1/data
image name:	exp1_ins_ssad_###.img
exposure time/image:	2.6 sec

Based on the chemical composition of the insulin crystals and the tabulated anomalous scattering lengths, the expected Bijvoet ratio as a function of the data collection wavelength can be calculated (Figure 3). At the data collection wavelength chosen, it is about 1.7%
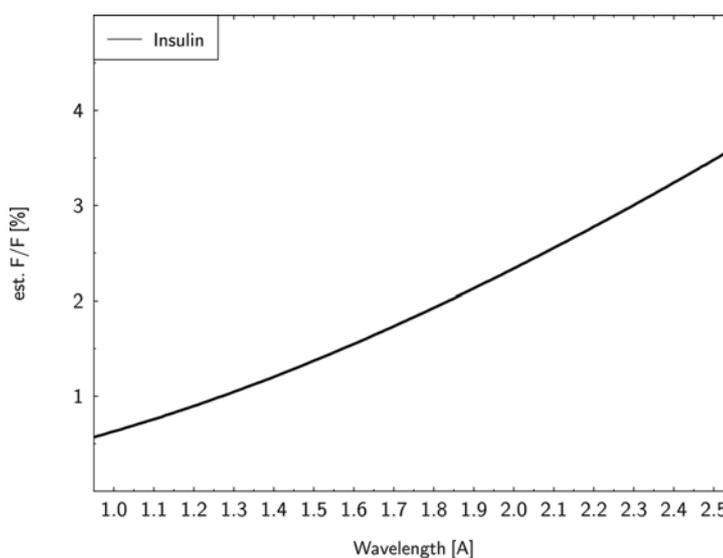
Figure 3: Estimated ΔF/F for Insulin as a function of data collection wavelength. The chemical composition used was $C_{255} H_{376} N_{65} O_{75} S_6$.
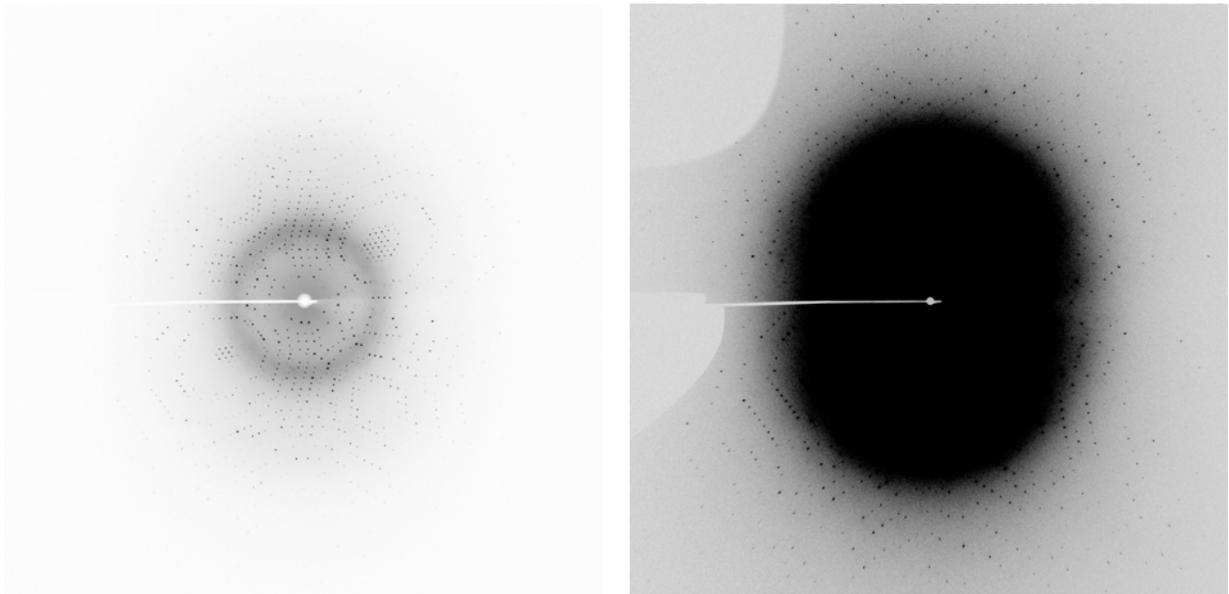
Figure 4: Diffraction image of cubic insulin displayed at different contrast levels. The shadow in the upper left hand corner on the image on the right originates from the cryo-nozzle. The other shadow on the left side is caused by the beam stop and its holder.

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 insulin, the correct space group is $I2_13$ (space group number 199) with unit cell parameters a = 78.0 Å.

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). The numbers in parentheses refer to the outermost resolution limit.

Resolution limits [Å]	50.0-1.60 (1.70-1.60)
Space group	$I2_13$
Unit cell parameters a, b, c [Å]	78.0
Mosaicity [˚]	0.14
Total number of reflections	208,859
Unique reflections	20,226
Redundancy	10.3 (8.3)
Completeness [%]	99.9 (100.0)
I/σ(I)	48.1 (13.6)
$R_{r.i.m.}$ / R_{meas} [%]	3.7 (17.0)
Wilson B-factor	22.1

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=ins_ssad.hkl CCP4

INPUT_FILE=ins_ssad.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 ins_ssad_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 can be solved using the SAD-protocol (run in the advanced version) 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). In the following the automatically generated summary of AUTO-RICKSHAW is printed together with the results of the structure determination:

The input diffraction data (file XDS_ASCII.HKL) were uploaded and then prepared and converted for use in Auto-Rickshaw using programs of the CCP4-suite (CCP4, 1994). ΔF-values were calculated using the program SHELXC (Sheldrick *et al.*, 2001; Sheldrick, 2008). Based on an initial analysis of the data, the maximum resolution for substructure determination and initial phase calculation was set to 1.8 Å. All of the six heavy atoms requested were found using the program SHELXD (Schneider & Sheldrick, 2002) with correlation coefficients CC(All) and CC(weak) of 53.3 and 32.2, respectively, and with a clear drop in occupancy after site no. 6 (Figure 5). This indicates that the correct solution was most likely found. The following figure shows the PDB coordinates of the six S-atom sites identified by SHELXD after occupancy refinement and in Figure 6 the six S-atoms superimposed on the anomalous difference electron density map are displayed.

```
CRYST1   78.000   78.000   78.000  90.00  90.00   90.00
SCALE1      0.012821  0.000000  0.000000        0.00000
SCALE2      0.000000  0.012821  0.000000        0.00000
SCALE3      0.000000  0.000000  0.012821        0.00000
HETATM   1  S   HAT     1      32.925  25.028  17.930 1.000 20.00
HETATM   2  S   HAT     2      31.772  26.363  16.489 0.969 20.00
HETATM   3  S   HAT     3      38.900  32.547  21.620 0.946 20.00
HETATM   4  S   HAT     4      40.450  32.487  20.035 0.889 20.00
HETATM   5  S   HAT     5      46.985  30.399  26.264 0.819 20.00
HETATM   6  S   HAT     6      48.007  30.424  24.246 0.813 20.00
HETATM   7  S   HAT     7      21.721  21.721  21.721 0.062 20.00
HETATM   8  S   HAT     8      38.768  34.806  19.971 0.184 20.00
END
```

**Figure 5:** PDB format output of the heavy atom sites found by SHELXD (Schneider & Sheldrick, 2002)

Figure 6: Anomalous difference Fourier electron density map with the six heavy atoms sites from SHELXD. The map is contoured at 8 σ.

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 and extended to 1.60 Å resolution. 90.2% of the model was built using the program ARP/wARP 7.0 (Perrakis *et al.*, 1999; Morris *et al.*, 2002). More details can be found in the attached AUTO-RICKSHAW output (directory exp1/struct_sol). The complete Auto-Rickshaw run in the advanced version took around 20 minutes. The model was then further modified using COOT (Emsley, 2004) and refined using REFMAC5 (Murshudov *et al.*, 1997). Figures 7 and 8 show snapshots of the final model superimposed with the anomalous difference map and the experimental electron density map after density modification in DM.
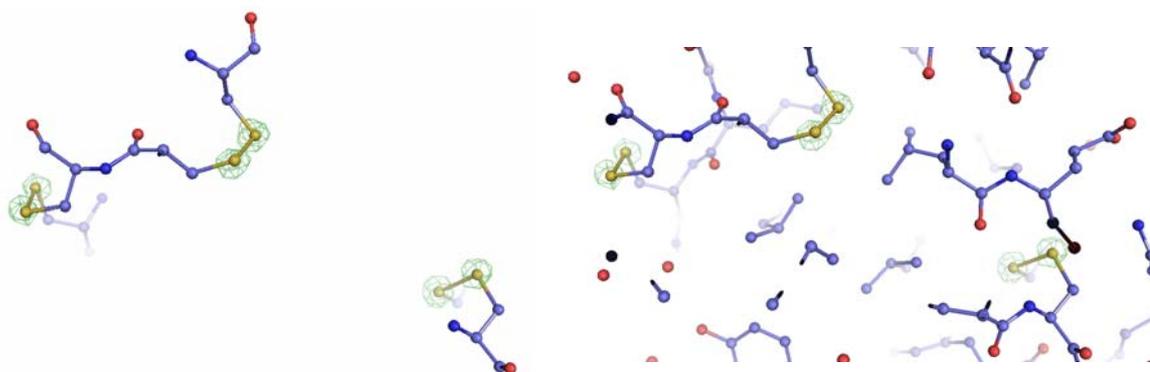
Figure 7: Final model superimposed with the anomalous difference Fourier electron density map. Left panel: the disulfide bridges in the final model; Right panel: a larger part of the final model. The map is contoured at 8 σ.
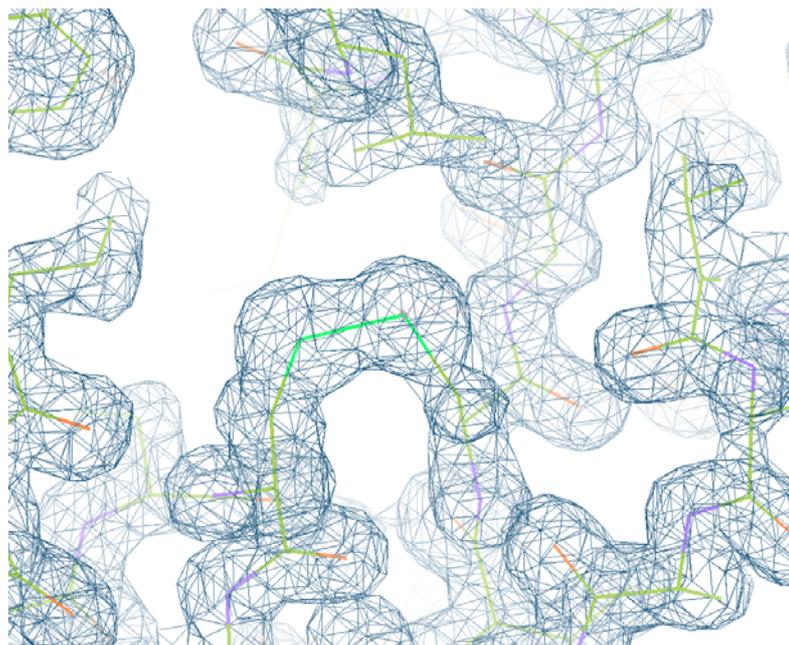
Figure 8: Experimental electron density map after solvent flattening using the program DM superimposed onto the final refined model. The map is contoured at 1.5 σ.

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.

Dauter, Z., Dauter, M., de La Fortelle, E., Bricogne, G. & Sheldrick, G. M. (1999). *J. Mol. Biol.* **289**, 83-92.

Djinovic Carugo, K., Helliwell, J. R., Stuhrmann, H. & Weiss, M. S. (2005). *J. Synch. Rad.* **12**, 410-419.

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

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

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

Hendrickson, W. A. & Teeter, M. M. (1981). *Nature* **290**, 107-113.

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.

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

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

Murshudov, G. N., Vagin, A. A. and Dodson, E. J. (1997). *Acta Cryst.* **D53**, 240-255.

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.

Voet, D., Voet, J. & Pratt, C. W. (2006). Fundamentals in Biochemistry - Life at the molecular level, 2nd Edition, John Wiley & Sons, Inc., Hoboken, NJ, USA.

Weiss, M. S., Sicker, T. & Hilgenfeld, R. (2001). *Structure* **9**, 771-777.