

# Basic SHELXC/D/E tutorial

Andrea Thorn

## Overview

In this tutorial, SHELX will be used for experimental phasing. Two cases will be processed: An easy S-SAD case and a MAD case. Several additional tasks are given in the end. We will use the command line and COOT for this. I am very grateful to Tim Gruene and George Sheldrick for their kind advice!

An online version of this tutorial hand-out can be found at: <http://shelx.uni-ac.gwdg.de/~athorn>

## Program availability

SHELX is available from <http://shelx.uni-ac.gwdg.de/SHELX/> and free for academic use.

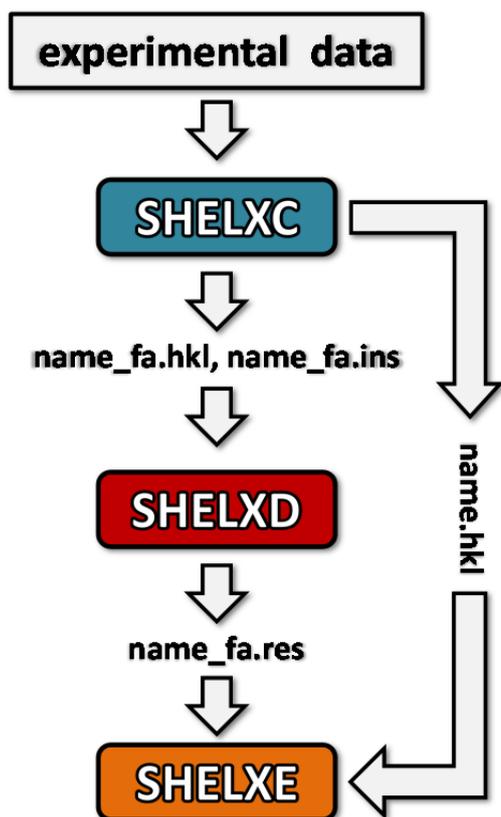
SHELX beta versions can be obtained from Prof. George Sheldrick, [gsheldr@shelx.uni-ac.gwdg.de](mailto:gsheldr@shelx.uni-ac.gwdg.de)

MAC versions are only compatible with an Intel CPU and 64-bit OS, not with PPC or 32 bit.

## Installation

To install SHELXC/D/E on your computer put the executable files into a directory in the 'PATH'.

## General data flow



SHELXC sets up files and calculates the estimated substructure structure factor amplitude  $|F_A|$  and the phase shift  $\alpha$  from input data.

<b>name_fa.ins</b>	cell, symmetry and instructions for SHELXD
<b>name_fa.hkl</b>	$h, k, l,  F_A , \sigma( F_A )$ and the estimated $\alpha$ angles
<b>name.hkl</b>	$h, k, l,  F_{obs} ^2, \sigma( F_{obs} ^2)$

SHELXD finds the marker atom substructure.

<b>name_fa.res</b>	potential marker atom positions
--------------------	---------------------------------

SHELXE applies density modification to establish the correct handedness, if necessary, and can autotrace the poly-Ala backbone given good resolution native data.

<b>name.phs</b>	contains information for map generation
<b>name.hat</b>	improved marker atom positions
<b>name.pdb</b>	if auto-tracing has been used
<b>name.pha</b>	phases for marker atom substructure

# S-SAD Phasing

<b>Molecule:</b>	<b>Viscotoxin-A1</b>
<b>Number of amino acids</b>	92
<b>Resolution</b>	1.25 Å
<b>Space group</b>	P4 <sub>3</sub> 2 <sub>1</sub> 2
<b>Cell</b>	a=b=65.73Å c=47.16Å α=90° β=90° γ=90°
<b>Strategy</b>	In-house CuK <sub>α</sub> (λ=1.54178 Å) for S-SAD

Pal A, Debreczeni JE, Sevvana M, Gruene T, Kahle B, Zeeck A, Sheldrick GM. *Acta Crystallogr. D64* (2008), 985-92.

These data were collected by Aritra Pal on a Cu home source. The solution of this structure is also discussed in Bernhard Rupp's book "Biomolecular Crystallography", Garland Science 2010. pp. 530-535.

## Data preparation with SHELXC

We start in the directory which contains the data, which is named `visc-a1` or something similar. The data are contained in the file `anom_data.hkl`. SHELXC can read data from XDS (`XDS_ASCII.HKL`), `sca` files and SHELX `hkl` files. It is best to read in `XDS_ASCII.HKL` directly rather than processing it with other programs, so that the data are *unmerged*, which is very desirable. You should create a new file, which could have really any name (and file extension); here it is '`visc-a1.txt`'. It has to be in text format and should contain the following lines:

```
CELL 65.73 65.73 47.16 90 90 90
SPAG P43212
FIND 12
NTRY 100
SAD anom_data.hkl
SFAC S
```

Cell parameters  
Space group  
Number of marker atoms\*  
Number of tries (100 is sufficient in this case)  
Filename of data set  
Marker atom type; only for naming

\* NOTE: Disulfide bridges are 2.03 Å long, so if we cut the resolution to 2.1 or less, we have to change this number: The disulfide peaks should then fuse into one anomalous density peak!

```
shelxc visc-a1 < visc-a1.txt
```

Note that the first '`visc-a1`' is the project name you want to assign and is used to generate the names of the files that are written by SHELXC. The second is the name of the input file. Having the same name makes it easier to find the input file later, but is not required!

Look at your screen output! There is a table like this:

Resl.	Inf.	9.11	5.61	4.22	3.45	2.95	2.60	2.33	2.12	1.96	1.82	1.70
N(data)		143	288	476	669	867	1047	1296	1490	1592	1880	2096
Chi-sq		1.09	0.94	0.91	0.93	0.89	0.94	1.02	1.13	1.14	1.15	0.96
<I/sig>		88.4	125.0	138.2	130.7	105.6	72.6	58.5	39.1	19.5	8.0	3.8
%Complete		127.7	98.6	99.0	99.4	99.7	99.6	99.8	99.7	99.7	99.8	99.2
<d"/sig>		1.76	3.50	2.41	2.01	1.52	1.35	1.25	1.12	0.91	0.81	0.74
CC(1/2)		89.0	86.7	77.3	61.2	48.2	37.1	29.7	22.6	6.0	-1.1	-1.9

It gives against resolution:

- The number of reflections read in
- $\chi^2$  which should be about 1 if the agreement of equivalents agrees with their estimated standard deviations.
- The average intensity divided by its standard deviation  $\sigma$
- Completeness (in %)

- $d''$  divided its estimated standard deviation  $\sigma$ , giving you a good indication for the strength of the anomalous signal, which should asymptote to 0.8 in the outer shell, if the data are processed well. If this line is missing, you should check if you gave an input data set in which the Friedel opposites have been merged.
- The self-correlation coefficient for the anomalous signal. It should be above 25% for a significant signal.

In SHELXD, the anomalous differences will be used to find the marker atom positions. Where should you cut the data for SHELXD? This value should be varied for more difficult cases, as it is often crucial!

Output files from SHELXC are:

**visc-a1\_fa.hkl** contains  $||F_{hkl}| - |F_{-h-k-l}||$  (Bijvoet differences) and estimates for the angles  $\alpha$   
**visc-a1\_fa.ins** contains instructions for SHELXD  
**visc-a1.hkl** contains intensity data for SHELXE

You can look at these files with a text editor. Where is the resolution cut-off for SHELXD?

## Finding the marker atom substructure with SHELXD

Just run from the command line:

```
shelxd visc-a1_fa
```

Output files from SHELXD are:

**visc-a1\_fa.res** marker atom solution with the best CFOM; this file is updated while SHELXC runs  
**visc-a1\_fa.lst** log file

### Did it work?

The **best CFOM value** gives an initial indicator on solutions. It should be at least over 35 a good S-SAD solution. For a good solution,  $CC_{weak}$ ,  $CC_{all}$  and CFOM should be high!

Have a look to the file visc-a1\_fa.res! What are the different columns?

SHELXD always tries to find more marker atoms than requested in the FIND instruction. This is why you can get more sites in the output. A sharp drop in occupancy between sites can indicate a good solution. However, it can be difficult to discern correct solutions at this stage. You should process the best solution (in terms of CFOM) to SHELXE.

How could you get even more high-occupancy marker atom positions than you expected in S-SAD?

## Density modification and auto-tracing with SHELXE

SHELXE carries out density modification and auto-tracing of the protein backbone. With these good data, all other parameters are sufficient at default values. Even the solvent content defaults to 45%, but here we expect 50%.

(The volume of the rectangular unit cell is approximately  $66 \text{ \AA} \times 66 \text{ \AA} \times 47 \text{ \AA} = 204732 \text{ \AA}^3$ . There are 8 asymmetric units per unit cell, which can be looked up in the International Tables A. An average amino acid in a crystal occupies  $140 \text{ \AA}^3$  of space. We expect two chains à 46 amino acids in the asymmetric unit, 92 in total. They would hence occupy  $92 \times 140 \text{ \AA}^3 \times 8 = 103040 \text{ \AA}^3$ .  $103040 \text{ \AA}^3$  divided by  $204732 \text{ \AA}^3$  is approximately 0.50, which can then be the protein content.  $1 - 0.5$  is 0.5 which is the solvent content.)

Please use these two commands in separate windows:

```
shelxe visc-a1 visc-a1_fa -a5 -h -s0.5
shelxe visc-a1 visc-a1_fa -a5 -h -s0.5 -i
```

This will read in

**visc-a1.hkl** native data

**visc-a1\_fa.hkl** anomalous differences,  $\alpha$   
**visc-a1\_fa.res** marker atom substructure

At this stage, we cannot know if the substructure or its mirror image is the correct solution. The **-i** option ensures that both enantiomorphs are tested.

NOTE: SHELXE input follows the general form  
**shelxe XX YY <options>**

XX is the native data set

YY is the experimental data file prepared by SHELXC or XPREP.

The output files are:

**visc-a1.lst** Log file  
**visc-a1.pha** Anomalous density map  
**visc-a1.pdb** Backbone trace with 92 traced residues (out of 92!)  
**visc-a1.hat** Improved substructure coordinates (hat stands for heavy atom solution)  
**visc-a1.phs** hkl file with phases to display electron density

Running SHELXE will take some time. Please take care to study the output! What information is in there?

### Selected SHELXE options

- h** Use this option if the marker atoms are part of the native structure and data. If their number is known, append it to h, e.g. **-h12**.
- a[N]** Do N cycles of auto-tracing of the protein's backbone.
- q** Explicitly search for helices.
- i** Invert the enantiomorph of the substructure!
- s** Solvent content, set to 45% by default.
- e** free lunch, e.g. **-e1.2** (particularly useful when data are missing, or when auto-tracing is not feasible – low resolution and nucleic acids)
- z** optimize marker atom substructure before starting density modification

## Did it work?

If the auto-tracing actually worked can be judged by the '**CC for partial structure against native data**' in SHELXE. In several rounds of auto-tracing, it should vary at a low value, increase suddenly and then vary at a higher value, which is higher than 25% for a successful trace at resolutions up to 2.2Å. Here, it should rise over 25% in the fourth cycle of auto-tracing.

**How do we discern the correct hand after running SHELXE?** Look at the values for the CC for the partial structure against native data. If you did not autotrace the structure, the correct hand is the one with the higher contrast which is printed to the screen at each cycle of density modification. This contrast can also be given graphically by HKL2MAP, as well as the less reliable map correlation coefficient. If in doubt: The correct map should look less ragged after some cycles of density modification. Clear side chains protruding from a traced backbone are an unambiguous sign of right hand!

You should use **COOT** to look at your results: Open the PDB file in COOT and then open the pha file. Therefore you have to use the option 'Open MTZ, mmCIF, fcf or phs...'. You will be asked if the cell from the PDB file should be used for the map, and you confirm this. To see if the anomalous density confirms the positions from the hat (and PDB) file, heighten the sigma value for the map. You can also look at the autotraced structure in the PDB file or at the normal electron density in the phs file. Which elements/which residues may account for the marker positions? This density is a good, mostly model-bias free starting point for building a model as complete as possible before your first REFMAC refinement. It can be enhanced a little by having a free lunch in SHELXE.

# MAD

<b>Molecule:</b>	<b>Concanavalin A</b>
<b>Number of amino acids</b>	237
<b>Best resolution</b>	1.68 Å
<b>Space group</b>	I222
<b>Cell</b>	a=61.90Å b=87.06Å c=88.92Å $\alpha=90^\circ \beta=90^\circ \gamma=90^\circ$
<b>Strategy</b>	Three-wavelengths MAD data set

Data measured by Marianna Biadene and Ina Dix at DESY.

See also Hardman K.D., Ainsworth C.F. *Journal: Biochemistry* 11 (1972) 4910-4919

## Data preparation with SHELXC

The input text file should be given as:

```
CELL 61.90 87.06 88.92 90 90 90
SPAG I222
FIND 1
NTRY 100
PEAK conca_peak.sca
INFL conca_infl.sca
HREM conca_hrem.sca
```

Cell parameters  
Space group  
Marker atom number  
Number of tries (Good data, 100 sufficient)  
Filenames of data sets; **NAT filename.sca** could be a native data set with better resolution; LREM is also possible.

If you are using XDS format, just give the files from different integrations different names – SHELXC does not rely on the filename extension to determine the format!

```
shelxc conca_mn < yourinputfile.inp | tee shelxc.log
```

(The command 'tee' prints the output both to the terminal and to the file shelxc.log, so that the program output can be kept for further reference. '| tee shelxc.log' can be omitted if you run SHELX on a Windows machine.)

The last table in the log file gives correlation coefficients (%) between signed anomalous differences. As the data sets have to have consistent relative Bijvoet differences, this is an indicator of real anomalous signal or noise:

### Correlation coefficients (%) between signed anomalous differences

Resl.	Inf	- 8.0	- 6.0	- 5.0	- 4.0	- 3.5	- 3.0	- 2.8	- 2.6	- 2.4	- 2.2	- 1.96
HREM/PEAK	72.7	74.8	68.8	54.4	59.6	60.2	55.5	45.1	47.4	36.2	34.5	
HREM/INFL	72.4	74.3	67.1	48.9	60.0	55.8	49.7	42.8	38.6	29.6	34.5	
PEAK/INFL	94.5	91.3	90.6	84.8	84.6	79.0	70.7	63.5	53.9	50.8		

This correlation is so high that it is not necessary to truncate the data. If the data are of lower quality, they should be truncated where best correlation coefficient drops below 25%. To do so you have to change the **SHEL** line in the file name\_fa.ins correspondingly, which was written by SHELXC.

## SHELXD

```
shelxd conca_mn_fa
```

In the output, you can see that the best CFOM (which is  $CC_{All} + CC_{Weak}$ ) quickly rises above 60%, above one can be rather confident that the MAD solution in SHELXD is correct, so continue to SHELXE.

> **While SHELXC runs, try to estimate the solvent content.** (Hint: In I222, there are 8 ASUs per unit cell.)

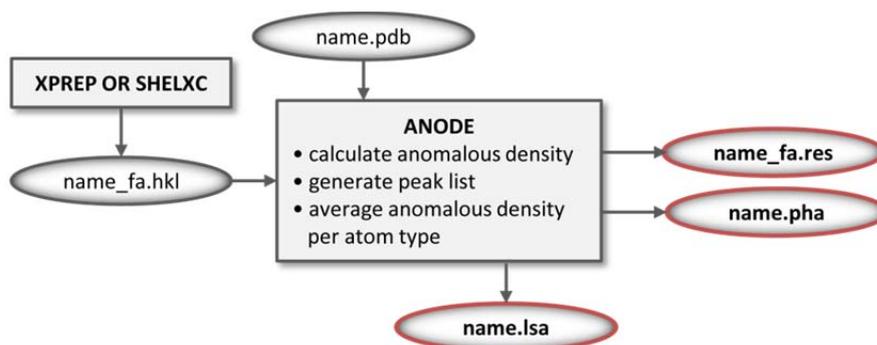
## SHELXE

```
shelxe conca_mn conca_mn_fa -a
shelxe conca_mn conca_mn_fa -a -i
```

Here, no other options are needed. (The solvent content is 36% which is close enough to the default.) Again, evaluate your results in COOT. Can you optimize the last SHELXE steps with other options? (A list of options is given on top of the output when SHELXE is run!)

## Extra tasks

- a) The folder co-conca contains Cu in-house data for SAD phasing of Concanavalin A with Cobalt. Try to solve the structure!
- b) Validation with anomalous density in ANODE: To validate the anomalous scatterer positions in your structure, and to find anomalously scattering ions in the hydration shell of your protein, you can use the tool ANODE. It will calculate the anomalous density for a PDB file.



Try it! In the folder visc-a1 is a final model of visc-a1 from the PDB and the file visc-a1\_fa.hkl you generated. Now rename 3C8P.pdb to visc-a1.pdb and give:

```
anode visc-a1
```

You get an output on screen and can read in vsc-a1.pha to COOT, just as the pha file from SHELXE, but now it's based on the phases from the final refined structure. Look at the sulfates.

- c) Try having a free lunch with Concanavalin A, expanding the data to 1.2 Å. Compare the maps in COOT. Which one would you take to start building a model, and why?
- d) Try solving both structures with HKL2MAP and make use of the graphical output of HKL2MAP!
- e) What happens if you give a better resolution by the SHEL command in SHELXD and also heighten the number of NTRY, for example NTRY 1000? Do the positions of the anomalous scatterers improve? Do solutions occur more frequently? (Hint: This is most comfortably done with the shelxd\_mp, the multi-processor version of SHELXD, as 1000 tries require some time.) Does this lead to better phases?
- f) Try the difficult SAD Tutorial!