CCP4 Workshop: Using `hkl2map` and `shelxc/d/e`

Tim Grüne

APS, June 2009

1 Setting Up

Testdata for this tutorial can be found at

`/home/data/shelx/sad`
`/home/data/shelx/mad`

Create a subdirectory for `hkl2map`:

```
mkdir hkl2map
cd hkl2map
```

2 `hkl2map` - SAD with Thermolysin

The program `hkl2map` is a graphical user interface for `shelxc/d/e` written by T. Pape and T. Schneider. Because the `shelx` programs are all script and command line driven and because people are often not used to this, `hkl2map` presents a very good starting point to get familiar with `shelxc/d/e`.
`hkl2map` can be used for SAD, MAD, SIR, and SIRAS. In addition to these types of experimant, `shelxc` can also do RIP phasing.

2.1 Preparation

Create a subdirectoy for the first data set, a thermolysin SAD experiment

```
mkdir tln
cd tln
cp /home/data/shelx/sad/tln_embo.sca
```

2.2 `shelxc` - Preparing input for `shelxd`

Start `hkl2map` from the command line simply by typing `hkl2map &` at the command prompt.
In the GUI at "Project Name" provide a word as an identifier for your work, *e.g.* `tln`, and type `ENTER`. The `shelxc` input mask opens.
Do not use spaces! It's best to stick to letters, numbers and underscores.
The default type of experiment is SAD, which is going to be used for this thermolysin data set.
There is no native data set so this field is left empty. If it were, usually at higher resolution, it could be entered here. It would be used be `shelxe` for density modification instead of the SAD dataset (`HA in` in the GUI).
Use the `Browse` button to load the data set `tln_embo.sca`. The fields for unit cell and space group are automatically filled in by the GUI, Fig. 2.2.
Notive that at this stage we cannot know, yet, whether the space group is $P6_122$ or $P6_522$. This question will get solved by the `shelxe` step further below.
You must confirm the spacegroup with the click-button before you can run `shelxc`.

2.2.1 Estimating the resolution cut-off for `shelxd`

Click on `View Graphics`. The first two graphs available under the `Display` menu show the general quality of the data. The third, $< d''/sig >$ allows to estimate the resolution cut-off that should be applied to `shelxd`.

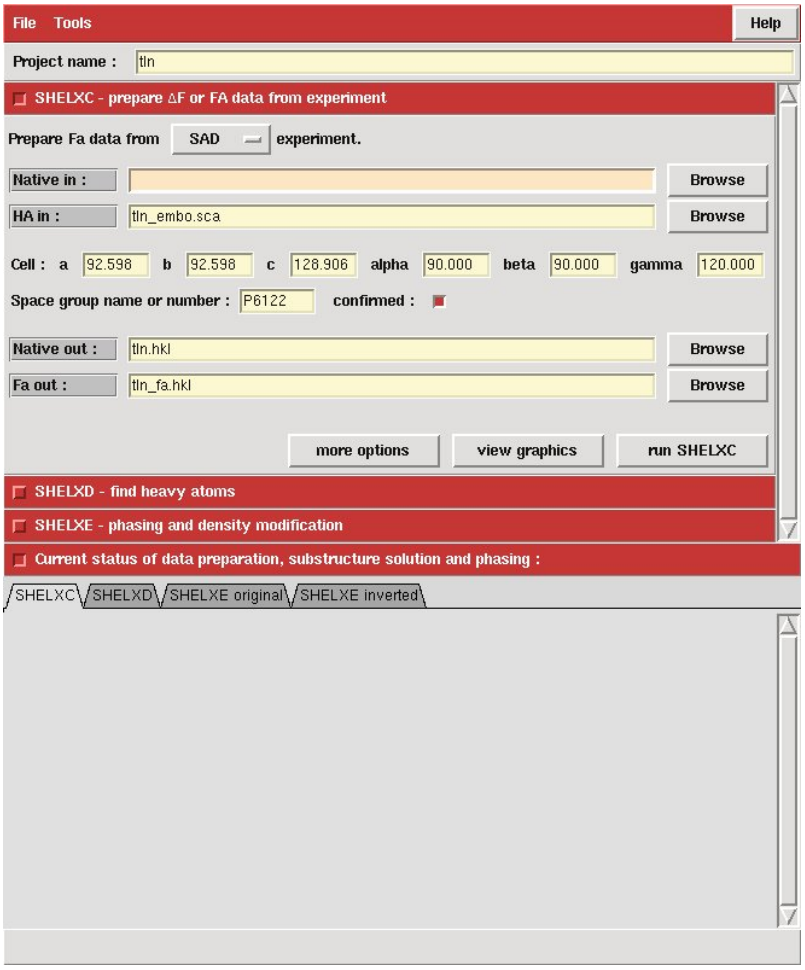

Figure 1: Main window of `hkl2map` after reading in the SAD data set.

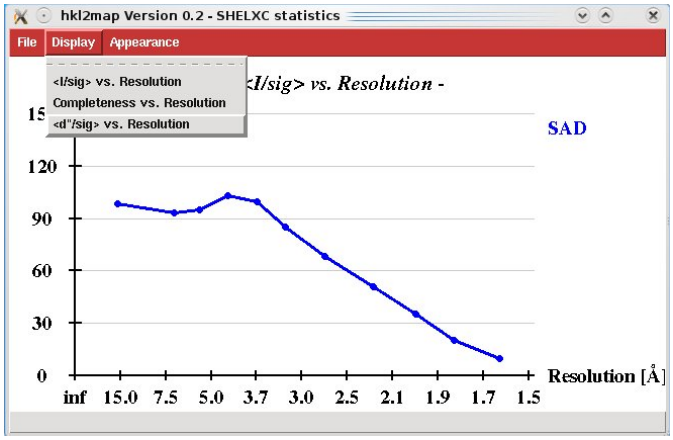

Figure 2: The $< d''/sig >$ plot from `shelxc`. The value is a strength of the anomalous signal in the data set. The data should be cut where the graph drops below 0.3.

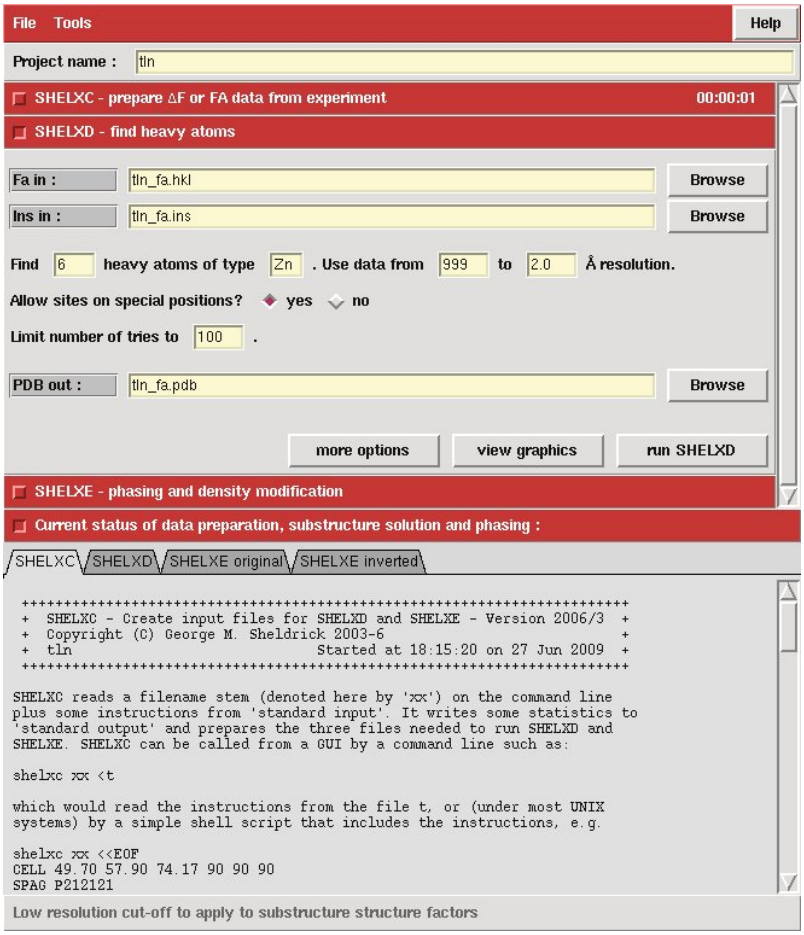

Figure 3: A sharp drop in the site occupancy usually indicates that a solution has been found. Values above 0.2 indicate real atoms contributing to the anomalous signal. Here, there are 1 Zn atom and 4 Ca atoms in the data set!

2.3 `shelxd` — finding the substructure

Close the `shelxc`-input mask by clicking on the small square in the red title bar and open the `shelxd`-input mask by clicking the small square next to "SHELXD".
We need to provide `shelxd` with

1. Expected number of type of marker atoms. The expected number should be within 20% of the real number of expected marker atoms. For *e.g.* halide soaks this number is difficult to guess and it may be worth running `shelxd` with several different settings.

2. The resolution cut-off for the anomalous signal, as *e.g.* estimated from the `shelxc` output.

Fill in the fields with 6 Zn atoms and a resolution cut-off of 2.0Å(see Fig. 2.3). Unlike macromolecules, ions have no restriction to symmetry. Therefore a Zn atom can happen to sit on a symmetry axis, and the option to allow atoms on special positions should be switched on.
In the case of SeMet phasing it can be left off since a Se atom in a SeMet residue cannot sit on a special position.
The worse the resolution the trials may be necessary to find a correct solution. 10,000 is not an unusual number!
You can click on `View Graphics` while `shelxd` is still running. For each trial set of random marker atom positions, `shelxd` prints the CC value between E_{obs} and E_{calc} calculated both from all data (CC_{all}) and from 30% of reflections which were not used during the dual-space refinement (CC_{weak}). The latter has a similar meaning as the R_{free} in model refinement. For SAD, $CC_{all} > 30\%$ is a good indication of a correct solution, but beware that the worse the resolution, the higher the CC-values are irrespective of a correct or incorrect solution.
An important graph is "Site occupancy vs. Peak Number". Are sharp drop after the expected number of sites is almost always a sign for a correct solution.

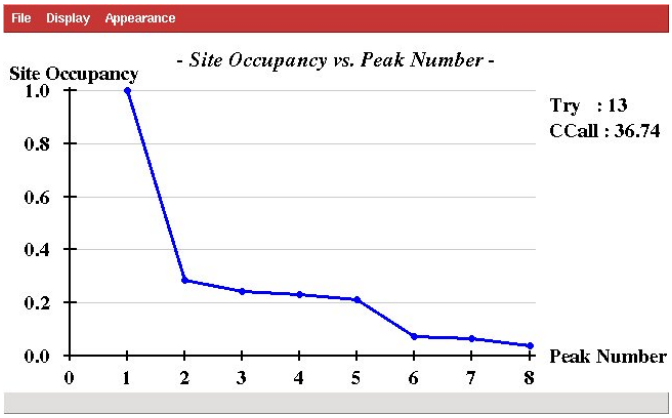

Figure 4: Input mask for `shelxd`

2.4 `shelxe` - Density Modification

The only additional input for `shelxe` is the solvent content. The extra button calculates this automatically from the number of residues in the asymmetric unit.
Thermolysin has one molecule with 316 residues in the asymmetric unit, *i.e.* a solvent content of 0.36 (Fig. 2.4)
The graph "Contrast *vs.* Cycles" (Fig. 2.4) distinguises between the correct and the wrong hand.
A lot better way to distinguish wrong from right solution is to look a the density map with *e.g.* Coot.
Start coot and load the Coordinate Files `tln.hat` and `tln_i.hat`. They contain the coordinates of the substructure, refined by `shelxe`, for the original hand (`tln.hat`) and the inverted hand (`tln_i.hat`). They are needed because the files created by `shelxe` that are used by coot to create the electron density maps do not contain the cell or symmetry information.
Then load the file `tln.phs`, the data for the electron density map for the original hand. When prompted for, select the cell with the space group $P6_122$. The map from the inverted hand can be calculated from the file `tln_i.phs`. When loading it make sure to select the space group $P6_522$.
This is because when inverting the hand of the substructure coordinates also the screw axes become inverted, so $P6_122$ turns to $P6_522$.

3 `shelxc/d/e` - the scripting approach

The `hkl2map` GUI does not incorporate all features available in `shelxc/d/e`, especially not the **autotracing** available in `shelxe`.
Therefore and also because better fine tuning in difficult cases is possible, it is worth learning how to use `shelxc/d/e` from the command line.
`hkl2map` produces all required input files and leaves them in the directory.

3.1 `shelxc`

The instructions for `shelxc` are found in the file `shelxc.in`. From the command line you can use it to achieve the same what `hkl2map` already did for you (ignore the "MAXM" entry written by `hkl2map`, it controls how much memory `shelxc` is asking from your computer. We need, however, add a few more options:

```
shelxc tlnscript << eof | tee shelxc.log
SAD   tln_embo.sca
CELL 92.598 92.598 128.906 90.000 90.000 120.000
SPAG P6122
FIND 6
SFAC Zn
eof
```

Unlike `shelxd` and `shelxe`, `shelxc` does not automtically create a log-file. The UNIX-command 'tee' in the above command creates a log file `shelxc.log` while printing it to the screen at the same time. In order to not overwrite the output from `hkl2map`, choose another descriptor for the output, like `tlnscript` in the above example.

`shelxc` calculates and sets up the three input files required by `shelxd`/ `shelxe`:

1. **tlnscript_fa.hkl** contains the non-anomalous data for the substructure

$$H \quad K \quad L \quad F_A \quad \sigma(F_A) \quad \alpha$$

   which is used by `shelxd`.

2. **tlnscript_fa.ins** input script with instructions for SHELXD, including symmetry operators, cell, number of heavy atoms to look for, ...

3. **tlnscript.hkl** the experimental data in **HKLF4** format, which means: each line contains the entries

$$H \quad K \quad L \quad F^2_{obs} \quad \sigma(F^2_{obs})$$

   It is read by `shelxe`and can later also be used for refinement.

3.2 `shelxd`

After the preparation by `shelxc`, `shelxd` is simply run by typing

```
shelxd tlnscript_fa
```

`shelxd` automatically adds `.hkl` and `.ins` to find the data and the insturction file `tlnsctript_fa.ins`
Wait until shelxd has finished.

3.3 `shelxe` - autotracing of the peptide backbone

The current version of `hkl2map` does not give access to the autotracing ability of the demo version of `shelxe`. Therefore run

```
shelxe_demo tlnscript tlnsctrip_fa -s0.36 -h5 -a2 -m30 -e1 -l3 -b
shelxe_demo tlnscript tlnscript_fa -s0.36 -h5 -a2 -m30 -e1 -l3 -b -i
```

`tlnscript` read the "native" data from `tlnscript.hkl`

`tlnscript_fa` read the anomalous data from `tlnscript_fa.hkl` and the marker atom coordinates from `tlnscript_fa.res`

`-s0.36` solvent content.

`-h5` The native data does contain the marker atoms (so it's not a "real" native data set), and `shelxe` should only use the first 5 atoms in the `.res`-file, because the other hits are just noise (judging from the occupancy)

`-a2` do 2 cycles of autotracing. For low resolution data, the resulting model is not going to be very complete, and one single cycle (-a1) is sufficient. For good data, 3–5 cycles produce best results (this tutorials chose 2 to speed up a little).

`-m30` 30 cycles of "classical" density modification

`-e1` use the **free lunch algorithm**. With data better than 2Å, `shelxe` can be used to invent phases and data beyond the actual measured data. The resulting maps are usually better than without this option.

`-l3` this is just a technical option: the Thermolysin contains more than 2,000,000 reflections, which is the default amount of memory `shelxe` allocates.

`-b` use the improved phases to also improve the substructure coordinates. the improved coordinates are written to `tlnscript.hat`.

`-i` invert the substructure. The resulting files will have `_i` appended to their basename so that no files will et overwritten.

NB: Don't put spaces between options and the values, `shelxe` will complain if you type `-s 0.36` instead of `-s0.36`.

4 Data Conversion

`shelxc` only read `.sca`-files and `.hkl`-files, which are plain files with one line per Miller index.

4.1 XDS

One can use the program `xprep` from Bruker to convert `XDS_ASCII.HKL` to a `sca` file. `xprep` is the more powerul predecessor of `shelxc`.
Or, if you have access to the scaling program `sadabs` (Bruker), you can use `xds2sad` (from the SHELX homepage) to convert `XDS_ASCII.HKL` to a file suitable for `sadabs`.

4.2 Mosflm/Scala

There are several ways:

1. use the option `outut polished unmerged` in textttscala to write a `.sca` file
2. use the program `mtz2various` from the ccp4i
3. use the program `mtz2sca` which comes with the `shelx` programs.

5 Availability and Installation

5.1 `shelx`

The `shelx` programs are available through http://shelx.uni-ac.gwdg.de/SHELX
The programs run on various computer platforms and the source code is available.
It is free to academic users. After filling in and sending the fax form available from this web-site, you are going to receive an email with download instructions and installation instructions.

5.2 `hkl2map`

`hkl2map` is available via a web interface at http://webapps.embl-hamburg.de/hkl2map/ where one can register and receive download instructions.

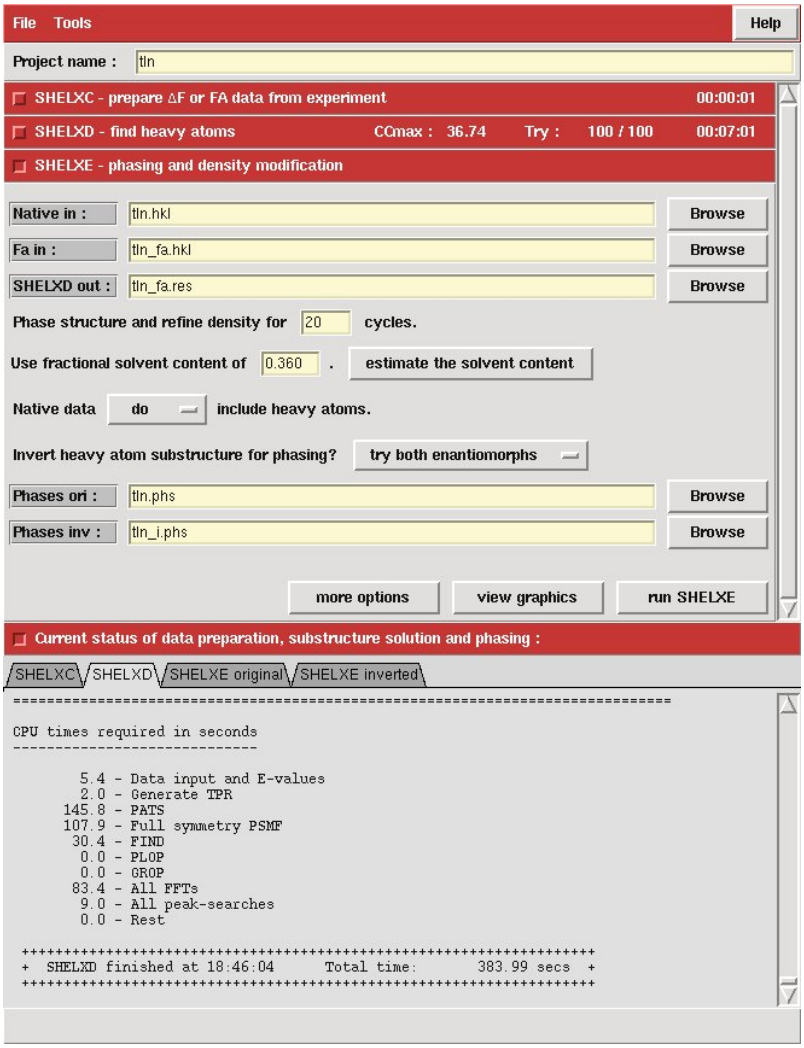

Figure 5: The solvent content of Thermolysin is about 0.36. The dafault of 20 cycles of density modification can be left untouched. In tricky cases, 100-200 cycles give better results.

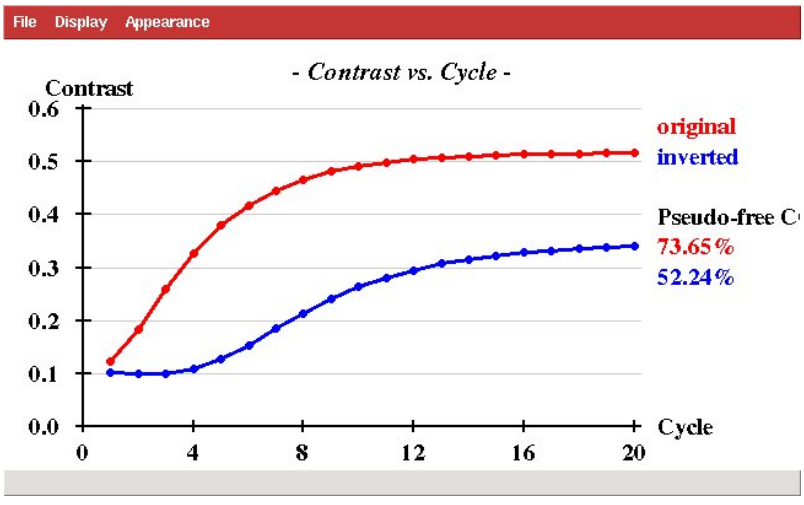

Figure 6: The electron density from the correct hand of the substructre creates a map with a strong contrast between solvent (strong variations and strong density regions) and non-solvent region (flat).