CCP4 Workshop Chicago 2011

Phasing with shelx c/d/e & hkl2map

Tim Grüne
Dept. of Structural Chemistry, University of Göttingen

June 2011

http://shelx.uni-ac.gwdg.de

tg@shelx.uni-ac.gwdg.de

Who is Who

shelxc, shelxd, shelxe [1] are three *command line* programs suitable for experimental phasing. They are available as part as the *SHELX-Suite* from http://shelx.uni-ac.gwdg.de/SHELX.
The β-version of shelxe with autotracing is available upon email request from George Sheldrick (gsheldr@shelx.uni-ac.gwdg.de)

hkl2map [2] is a *Graphical User Interface* for shelx c/d/e. It is very useful for making some of the critical decisions during the phasing process by its graphical output. It is available from http://webapps.embl-hamburg.de/hkl2map

shelxc

shelxc is a “pipeline”-version of the far more versatile program *xprep* (Bruker AXS).

1. Data preparation (merging of data sets, extraction of anomalous signal)
2. Data statistics (resolution cut-off for shelxd, selection of data sets)
3. Preparation of shelxd input script

Available Phasing Scenarios

- SAD (single wavelength anomalous dispersion)
- MAD (multi wavelength anomalous dispersion)
- SIR (single isomorphous replacement)
- SIRAS (SIR with anomalous scattering)
- RIP (radiation damage induced phasing)

Shelxc Example Run

shelxc is run from the command line with the syntax *

```
shelxc mymad < my_shelxc.input | tee shelxc.log
```

`mymad` sets the project name and determines the filenames used by shelxd and shelxe. It cannot contain spaces or a period ".".

The names `my_shelxc.input` and `shelxc.log` are completely arbitrary.

E.g. for a MAD experiment, the text file `my_shelxc.input` may look like

```
NAT jia_nat.hkl
HREM jia_hrem.sca
PEAK jia_peak.sca
INFL jia_infl.sca
CELL 96.00 120.00 166.13 90 90 90
SPAG C2221
FIND 8
SFAC SE
```

*The construct `|tee shelxc.log` saves the output from shelxc in the file shelxc.log for later reference.

Shelxc Keywords

Depending on the experiment, the following keywords tell shelxc how to treat the data (keywords in brackets are optional):

SAD
- `SAD`
- `(NAT)`

SIRAS
- `SIRA`
- `NAT`

SIR
- `SIR`
- `NAT`

MAD
- `(NAT)`

at least 2 of
- `PEAK`
- `INFL`
- `LREM`
- `HREM`

RIP
- `BEFORE`
- `AFTER`
- `(NAT)`

Each keyword takes the filename of the corresponding integrated dataset, either in scalepack format (ending `.sca`) or hklf4-format (intensities, ending `.hkl`)

Shelxc: Wavelength

One very user friendly feature:

shelxc/d/e do not make use of wavelength information or of experimentally determined values for f'/ f''.

In borderline cases of MAD or SIRAS it may improve results, though, to take these into account.

In this case data have to be prepared with xprep which also allows to refine these values.

Shelxc Files

The call `shelxc mymad` creates three files. Their names are **not quite** arbitrary and should not be changed by novice users to avoid overwriting of the files and confusing error messages:

`mymad_fa.ins` Text file with instructions for shelxd

`mymad_fa.hkl` Artificial substructure data set from which shelxd determines substructure coordinates. Each line contains the estimate or calculated phase angle α. α is not used by shelxd, but by shelxe to calculate an initial phase estimate for the protein structure as

$$\phi_P(hkl) = \phi_S(hkl) + \alpha(hkl)$$

ϕ_S is the phase angle calculated from the substructure coordinates determined by shelxd.

`mymad.hkl` "native" data used by shelxe for phasing and density modification

Shelxc Output

Both shelxd and shelxd automatically create log-files (ending `.lst`).

shelxc only writes to the terminal it was started from (stdout in “unix’ish”). Therefore it is best “trapped” to a file with a program like ’`tee`’).

The output contains some useful analyses of the input data.

The GUI hkl2map plots graphs of these statistics and therefore the shelxc output will be discussed further below when hkl2map is introduced.

Shelxd — Finding the Substructure

`shelxd mymad_fa`

reads the "substructure data" `mymad_fa.hkl` and its instructions from `mymad_fa.ins`. The most important entries in `mymad_fa.ins`:

SFAC SE atom type to look for

FIND 12 expected number of substructure atoms, should be within 20 % of the actual number (try several for *e.g.* a soak where the number is not known)

SHEL 999 3.3 resolution limits of the anomalous signal, not the original data. High resolution limit can be critical, but the default of $d_{\text{max}} + 0.5$ Å works well in normal cases. Fine-tuning will be discussed in the hkl2map section

NTRY 1000 number of trials. Since shelxd starts from random atom positions, a low quality data set may require a large number of trials before a solution is found.

Shelxd Output

shelxd automatically writes a logfile `mymad_fa.lst`, so no redirection (as in the case of shelxc) necessary.

While shelxd is running, it writes the currently best substructure solution to `mymad_fa.res` which contains the substructure coordinates in fractional coordinates and which is later read by shelxe.

```
REM Best SHELXD solution:   CC 60.74   CC(weak) 49.22   CFOM 109.96
TITL mymad_fa.ins MAD in C2
CELL  0.98000  109.02   61.75   71.74   90.00   97.08   90.00
LATT  -7
SYMM -X, Y, -Z
SFAC SE
UNIT   192
SE01   1  0.758774  0.508636  0.246391   1.0000  0.2
SE02   1  0.792908  0.398262  0.138903   0.8845  0.2
                [...]
SE10   1  0.925819  0.231575  0.191291   0.5569  0.2
SE11   1  0.495239  0.183609  0.416278   0.5352  0.2
SE12   1  0.643097  0.029221  0.210653   0.4897  0.2 <---
SE13   1  0.811539  0.048553  0.227752   0.1453  0.2 <---
SE14   1  0.600281  0.156860  0.149628   0.0764  0.2
HKLF 3
END
```

The sixth column contains the occupancy of the corresponding atom. A sharp drop (here between SE12 and SE13) is a promising sign of a correct solution. In this case, shelxe could be run with the option `-h12` instead of just `-h`.
The correlation coefficient (CC and CCweak) in the first line measures the reliability of the solution

For SAD, a CC of more than 30 % is a safe sign of a correct solution, for MAD the limit is about 40 %.

Shelxe (1/2)

A short usage instruction of shelxe is printed by just typing `shelxe` at the command line. This will print an about 1 page usage instruction that explains various scenarios of how to use shelxe.

shelxe does not us an input file, all parameters are provided as command line options after native data and substructure solution.

A typical and one of the most simple command line could be

```
shelxe mymad mymad_fa -s0.65 -h12 -a -q
```

`mymad` read native/original data from `mymad.hkl`

`mymad_fa` read angle estimate for α from `mymad_fa.hkl`, substructure coordinates from `mymad_fa.res` (the shelxd output)

`-s0.65` Assume a solvent content of 65%. The solvent content is one of the most critical parameters for shelxe and it is worth testing various settings

`-h12` only use the top 12 positions from `mymad_fa.res` and assume that the original data contains the substructure coordinates. (*e.g.* in the case of SAD or MAD when PEAK is used as native data). `-h` without number uses all sites present in `mymad_fa.res`

Shelxe (2/2)

```
shelxe mymad mymad_fa -s0.65 -h12 -a -q
```

`-a` run 5 (default) cycles of poly-ALA autotracing. This feature is most useful and currently available in the β-version of shelxe (send an email to gsheldr@shelx.uni-ac.gwdg.de)

`-q` by default, shelxe searches for tri-peptides during the auto-tracing cycles. -q lets shelxe explicitly search for α-helices. Unless you know there are no helices (*e.g.* no protein at all), this option should always be used since it significantly improves the result.

Shelxe `-i`: Inverted Substructure

It is impossible to distinguish the substructure from its enantiomer with the anomalous data and there is a 50 % chance that the coordinates in `mymad_fa.res` are inverted w.r.t. the correct substructure.

Therefore shelxe must always be run two times

- with the direct hand
- with the inverted hand, *i.e.* with the same options as the direct hand *plus* the switch `-i`. This inverts the hand and takes care of everything necessary (*e.g.* inversion of screw axes, $P4_1$ to $P4_3$). The output files are amended by `_i` to distinguish the two runs.

N.B. if the inverted hand turns out to be the correct hand, your space group may change - usually in the presence of screw axes. Keep this in mind when you convert your native data to *e.g.* mtz-format!

Shelxe The Correct Hand

Criteria to distinguish correct from wrong hand:

1. Correct hand shows better Contrast, especially at early cycles of density modification.
2. Correct hand has higher map correlation coefficient throughout resolution range:

```
d     inf - 4.66 - 3.70 - 3.23 - 2.93 - 2.72 - 2.56 - 2.43 - 2.33 - 2.24 - 2.15
<mapCC> 0.626  0.795  0.775  0.754  0.819  0.804  0.756  0.694  0.620  0.582
<mapCC> 0.810  0.877  0.845  0.844  0.874  0.856  0.840  0.830  0.839  0.809
```

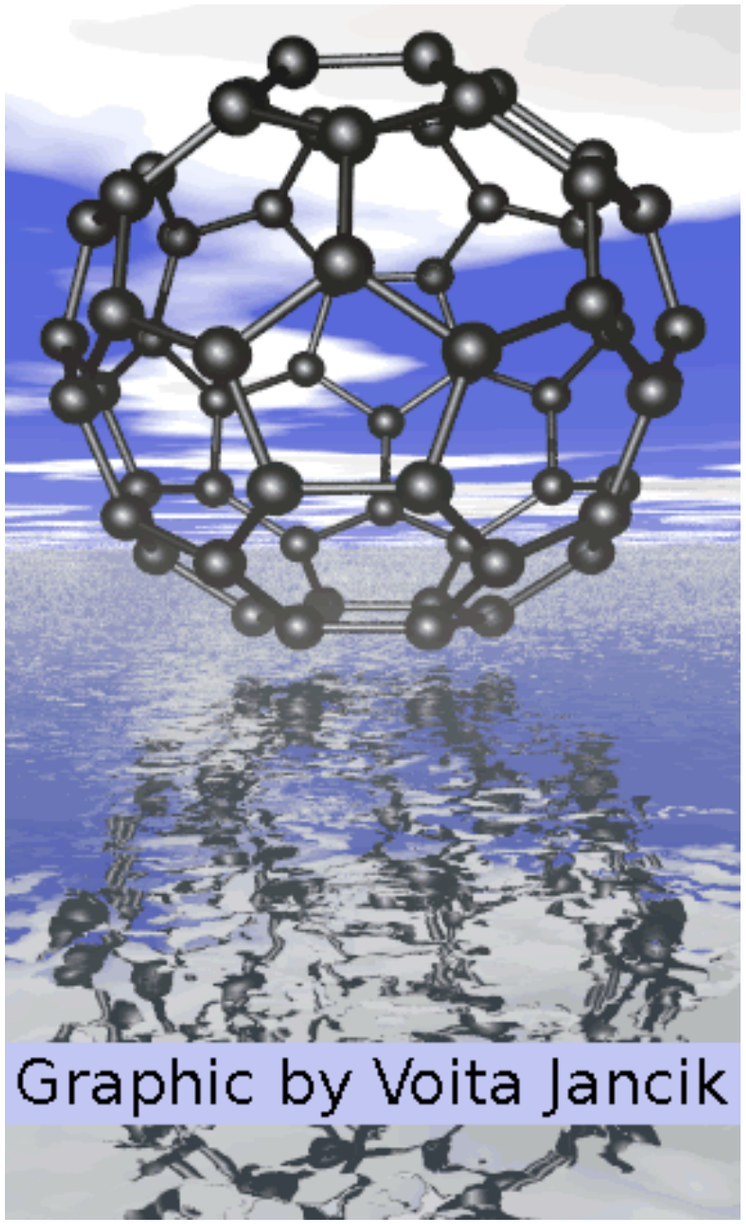
3. (with shelxe β-version, which you really should get a copy of) a better poly-ALA trace with at least 10 residues per chain

hkl2map plots the map contrast while shelxe is running and interrupts the wrong hand automatically.

The mapCC is plotted at the end of shelxe.

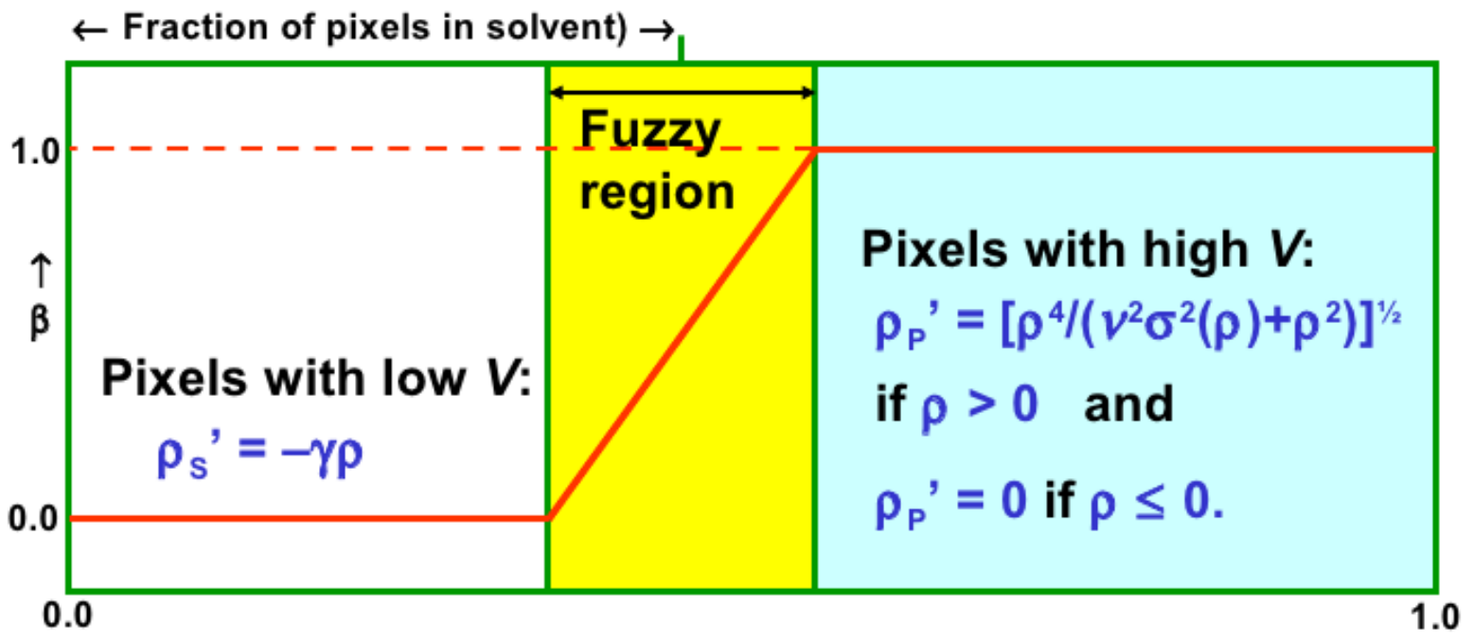
When using the auto-tracing option (`-a`) in shelxe, the first two figures (contrast/ mapCC) become rather meaningless, but in this case the poly-ALA trace tells unambiguously the correct hand.

Density Modification: The Sphere of Influence



Shelxe performs density modification based on the “sphere of influence” method.

- Calculate the variance of electron density at each map point on a sphere with 2.42 Å radius.
- 2.42 Å = typical 1-3-distance in macromolecules.
- regions with high variance: probably (ordered) protein region.
- regions with low variance: probably (disordered) solvent region.
- “sharpen” (enhance) protein region
- “flatten” solvent region.



The Sphere of Influence: The Solvent Content

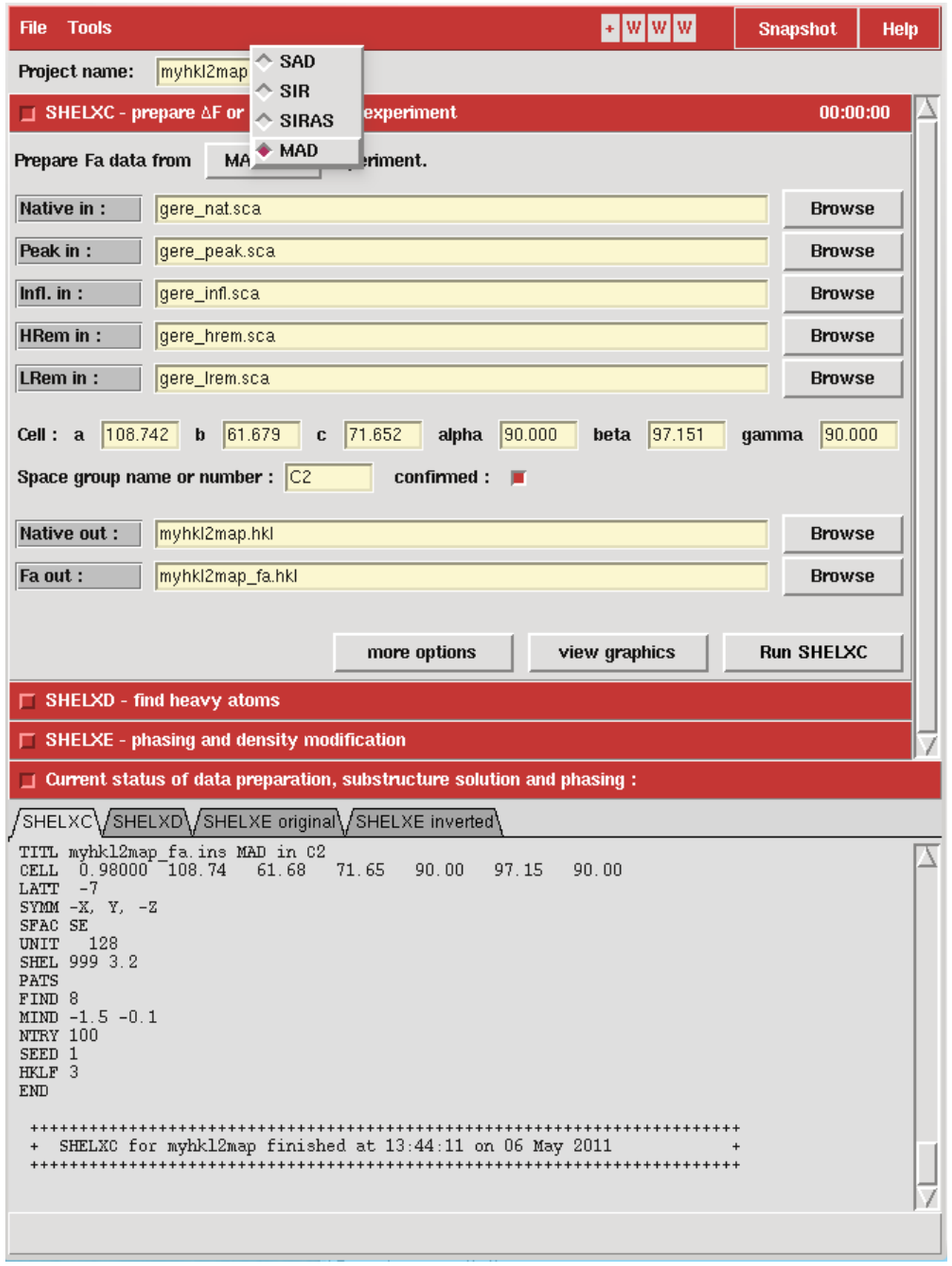
As a consequence of the “sphere of influence” algorithm shelxe is sensitive to the solvent content (option `-s`).

It can be calculated with `matthews_coef` (ccp4), hkl2map, or by hand assuming 140 Å^3/a.a.

At low resolution some parts of the molecule may be disordered and therefore *behave* like solvent. In such cases it may be worth assuming a higher solvent content than calculated.

Chain tracing in the current β-version of shelxe has reduced its sensitivity to the solvent content.

Hkl2map



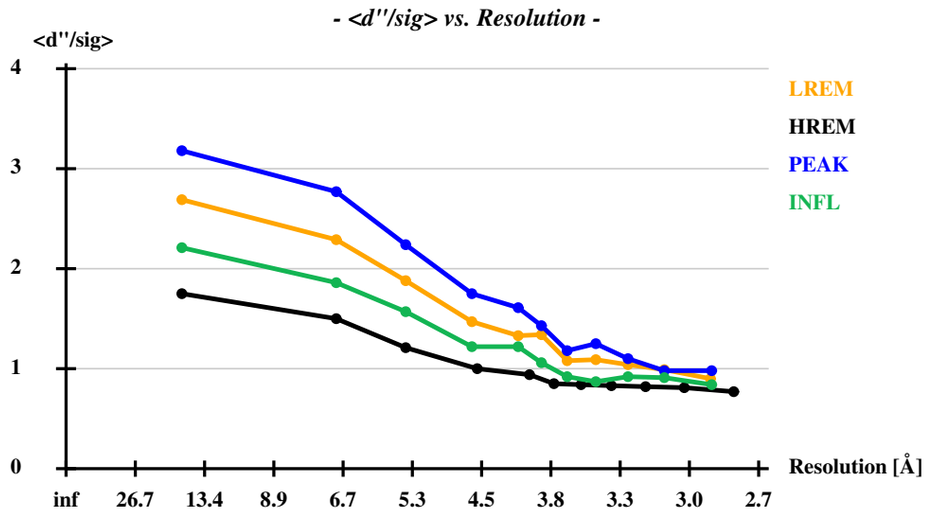
The hkl2map GUI unifies the three steps of shelxc, shelxd, and shelxe in one graphical user interface. It facilitates the input for SAD, SIR, SIRAS, and MAD (RIP is not supported).

The output filenames are determined by the “Project name” (as used as argument to shelxc — no spaces, no period “.”).

Its main advantage over running shelx c/d/e from the command line: the graphical display.

Hkl2map: Quality Control

Correctly estimated standard deviations of measured reflections are important for substructure solution - much more important than, *e.g.* for refinement.

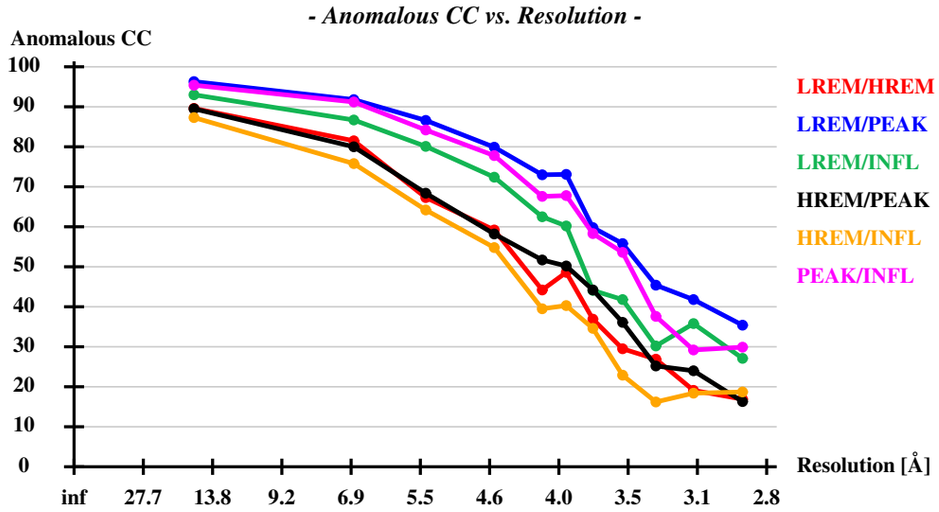


The plot `<d''/sig>` directly shows the strength of the anomalous signal.
As expected it is strongest for the `peak` data set.

The anomalous signal approaches 0.8 for random (non-anomalous) data. If the graph drops below this value, it is indicative of problems during data collection, *e.g.* improper background correction in the presence of ice rings (around 3–4 Å), or a poor error model.

Hkl2map: Quality Control

One can also check the “anomalous correlation coefficient” between data sets.



The correlation coefficient between data sets is an even better judge than the anomalous signal.
In the case of SAD, shelxc plots the “self-anom CC” (and so does pointless by P. Evans).
If one data set is much below the others it may be worth excluding that data set.

Hkl2map: Resolution Cut-off

Weak anomalous data contributes mostly noise. This can hamper the substructure solution.

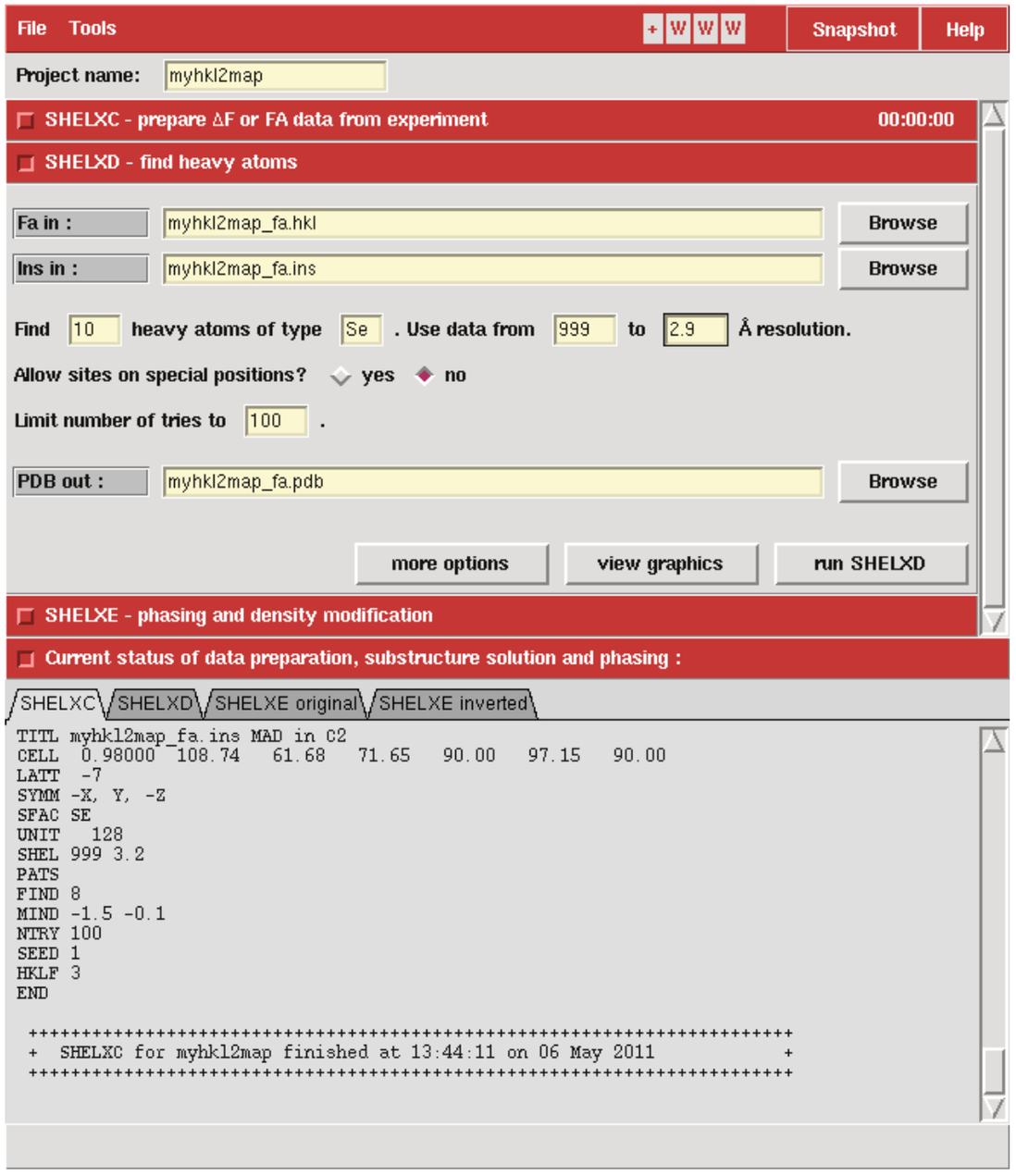
To avoid this, the resolution cut-off for the anomalous data can be set for shelxd with the `SHEL` keyword.

The two plots anomCC and $\langle d''/sig \rangle$ can be used to find a promising cut-off:

- include data up to $\langle d''/sig \rangle > 1.3$ or
- include data up to anomCC > 30 %

The default in shelxc is to add 0.5 Å to the high resolution limit of the native data set. This works sufficiently well for most cases.

Hkl2map — Starting Shelxd



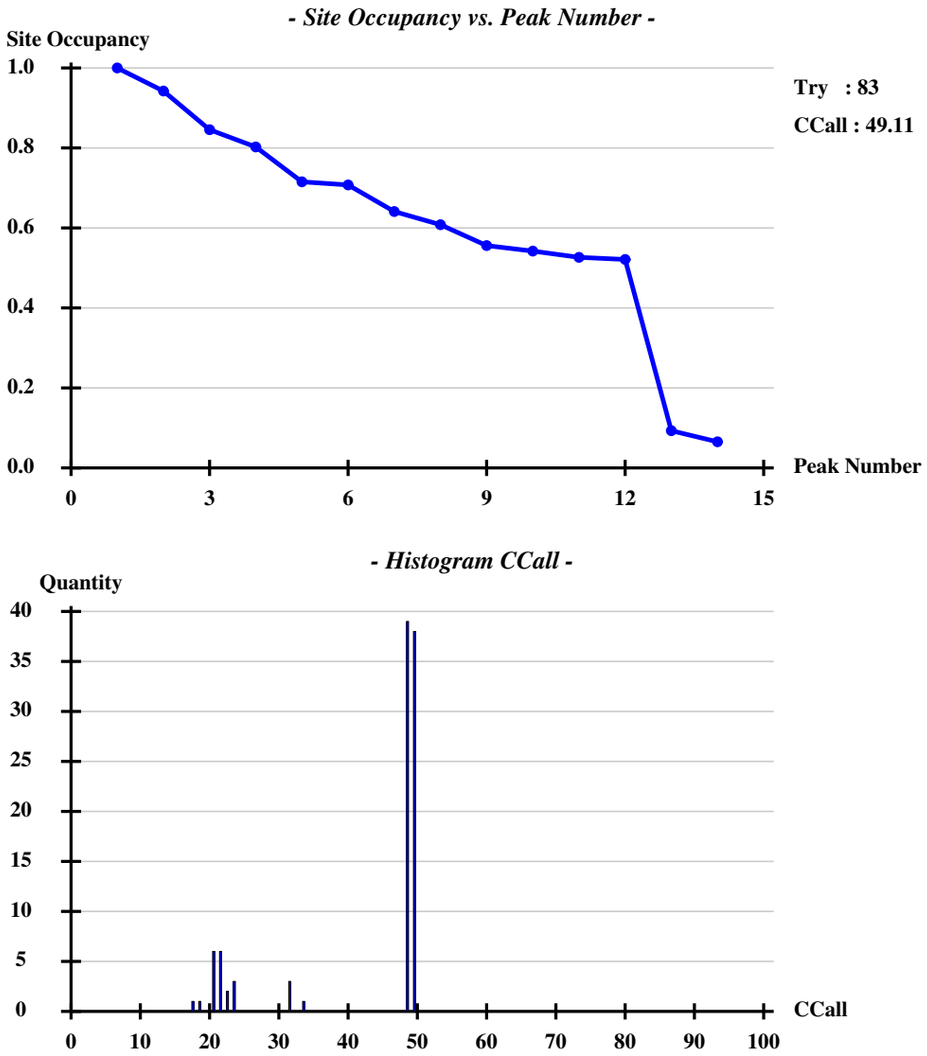
Enter the atom type and how many substructure atoms are expected.

Check the high resolution limit (the default $d_{\text{max}} + 0.5$ Å)

For halide or heavy atom soaks: allow sites on special positions.

The graphics are updated live while shelxd is running.

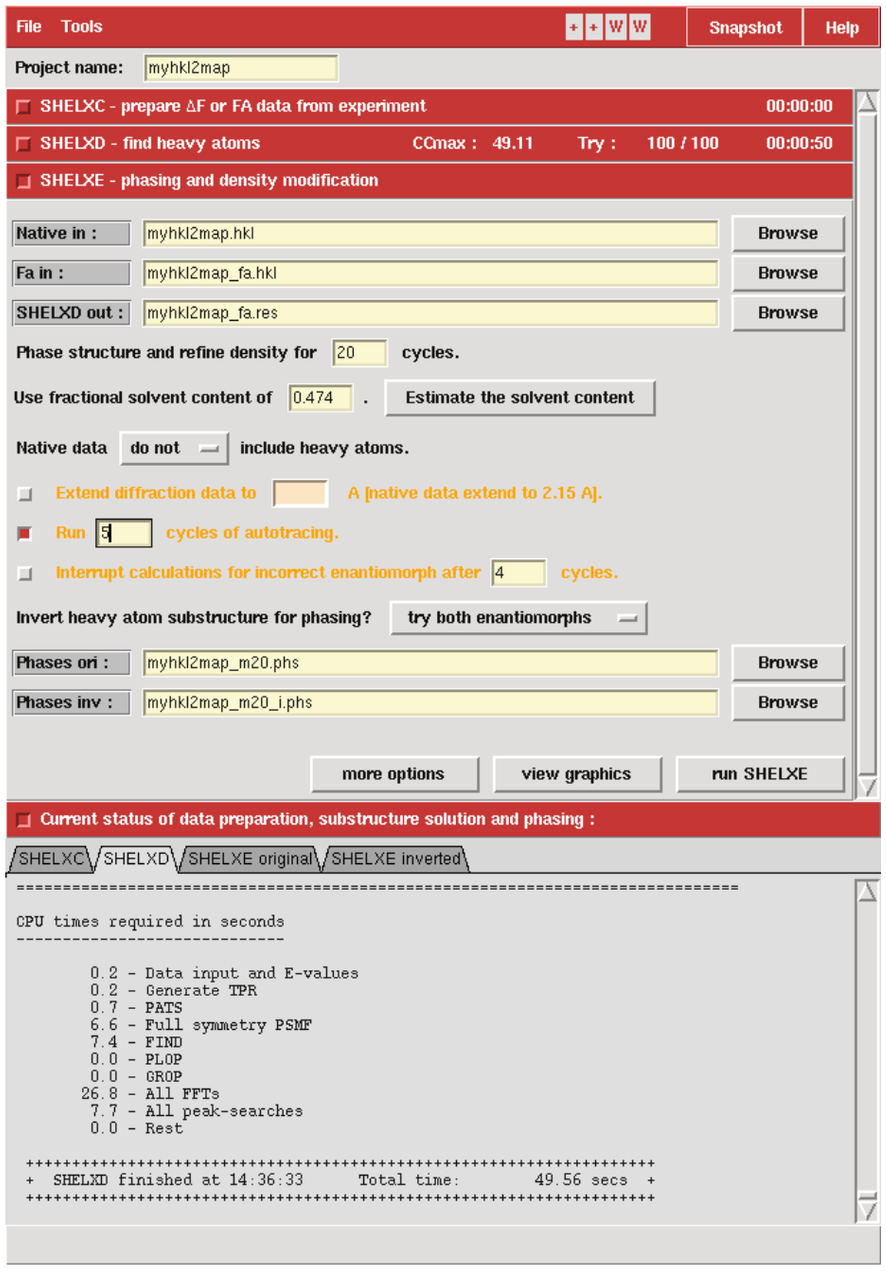
Hkl2map — Judging Shelxd



A sharp drop in the occupancy of the sites is a good sign.
Sites with an occupancy greater than 0.3 can be considered correct sites.
In difficult cases it is worth rerunning shelxd with the correct number of sites.

If one solution is found many times (high peak in histogram), this solution is most likely correct.
Another indicator for correctness: CC $>$ 40 % for MAD or $>$ 30 % for SAD and a not too low CCweak.
N.B. at low resolution (3.5 Å and worse) the CC automatically increases, be the solution correct or not).

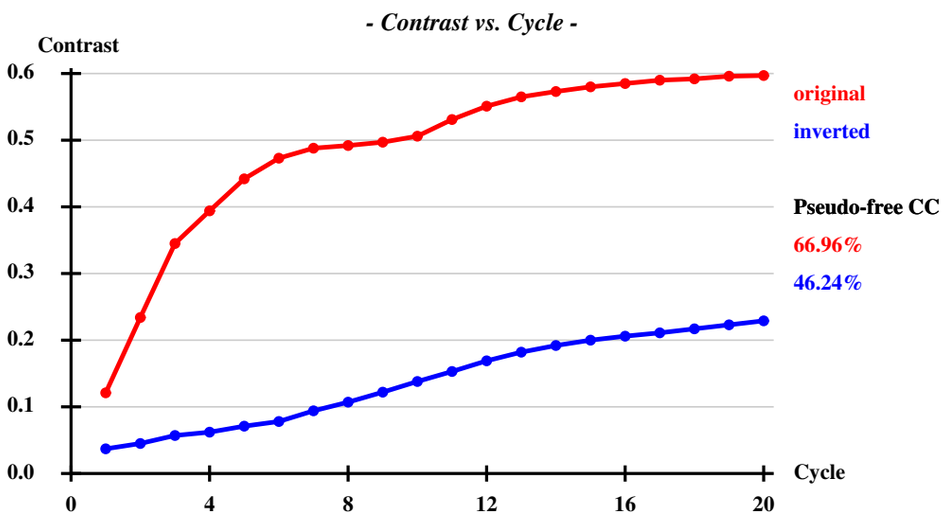
Hkl2map — Setting up Shelxe



The solvent content can be estimated from the number of residues in the asymmetric unit, assuming 140 $Å^3$/a.a.

The GUI can make use of the autotracing ability of shelxe. 5 cycles are usually sufficient.

Hkl2map — The Correct Hand



hkl2map runs shelxe both with and without the `-i` option.
The correct hand usually has a stronger map contrast especially at the beginning of density modification.

Life is much easier when autotracing works. The correct solution has an average chain length of $>$ 10 and a CC of $>$ 25 % (these number are fairly reliable)
By means of autotracing shelxe can produce an interpretable map even for borderline cases.

```
==> myhkl2map.pdb <==
TITLE myhkl2map.pdb Cycle 1 CC=41.49% 321 residues in 12 chains

==> myhkl2map_i.pdb <==
TITLE myhkl2map_i.pdb   Cycle 1 CC= 6.38% 103 residues in 12 chains
```

Shelxe Tweaks

Patience

In difficult cases one can increase the time shelxe spends searching for α-helices with the `-t` switch, *e.g.* `-t10` for a 10-fold longer search.

`-a10` carries out 10 cycles of autotracing instead of the default of 5 cycles.

In contrast to “only” density modification and to shelxd, autotracing is comparatively time consuming, and the above two switches elongate the duration even more, but the results are worth the waiting.

In the presence of NCS (non-crystallographic symmetry) one can supply `-n`. While this does not seem to improve results much, it is a good sign for the correct substructure solution if shelxe **does** find NCS in the substructure sites.

Shelxe: Substructure Recycling

shelxd writes the substructure coordinates to the `mymad_fa.res`-file.

shelxe writes the improved coordinates to `mymad.hat`. This file has the same format as the `.res`-file.

Recycling: Rename `mymad.hat` to `mymad_fa.res` and re-run shelxe. This provides shelxe with better starting phases and hence improves the results.

Caveat: If the inverted structure turns out to be the correct hand (*i.e.* `mymad_i.hat` from the `-i`-run of shelxe), the second run of shelxe must be run without the `-i` switch:

The `.hat` file is already corrected for the inverted hand.

Getting There . . .

File format conversion is a real pain in crystallography.

shelxc can read scalepack format and shelx hkl-files. These can be created depending on the integration software, *e.g.*

HKL2000 scalepack-format read directly

XDS 1. xds2sad + sadabs (only available from Bruker AXS)

2. xdsconv to create hkl- or sca-format. Note that you need "`MERGE=FALSE`" in `XDSVONV.INP` in order to create the self-anomalous CC plots
3. pointless + scala followed by mtz2sca [3]

`XDS_ASCII.HKL` can also be read by xprep, but than you do not need shelxc

Mosflm + Scala mtz2sca [3] converts to sca-format

. . . and Back

Coot reads substructure coordinates `.res` (shelxd) and `.hat` (shelxe)

Coot reads electron density map `.phs` (shelxe), but requires cell and symmetry, i.e. read in `.hat` or `.pdb` (from auto-tracing) first.

f2mtz (from ccp4) converts map `.phs` to mtz-file, e.g. for use with arp/warp [4]. phs-format: `H K L F FOM PHI SIGF`
Typical f2mtz-script (I call it `phs2mtz.sh`):

```
#!/bin/bash
f2mtz hklin mymad.phs hlkout mymad.mtz << eof
title My successful MAD data
pname EMBO Workshop 2011
dname shelxcde
cell  24.64 39.63 65.53 90 90 90
symm P212121
labout H K L Fmad FOMmad PHImad SIGFmad
CTYPOUT H H H F W P Q
END
eof
```

NB: The arp/warp GUI in ccp4i does not find the SIGF column automatically, it has to be given explicitly.

Other Shelx c/d/e Interfaces

Several other suites make use of shelx c/d/e. Two important ones:

Autorickshaw [6] pipeline for experimental macromolecular phasing which tries a large number of combinations of possible programs for substructure solution, density modification, and model building

Arcimboldo [5] computer intensive cluster for “ab initio” phasing even at 2 Å resolution: combines helix-search using phaser [8] with fragment extension by shelxe.
Arcimboldo requires native data to about 2 Å resolution and at least one α-helix, but no other phase information.

References

1. G. M. Sheldrick, *Experimental phasing with SHELXC/D/E: combining chain tracing with density modification*, Acta Crystallogr. (2010), D**66**
2. Pape T. & Schneider T.R ., *HKL2MAP: a graphical user interface for phasing with SHELX programs*. J. Appl. Cryst. (2004) 37
3. Grune, T., *mtz2sca and mtz2hkl: facilitated transition from CCP4 to the SHELX program suite* J. App. Cryst. (2008), 41(1)
4. Morris, R. J. and Perrakis, A. and Lamzin, V., *ARP/wARP's model–building algorithms. I. The main chain*, Acta Crystallogr. (2002), D58
5. Rodríguez, D. D. *et al.*, *Crystallographic* ab initio *protein structure solution below atomic resolution*, Nature Methods (2009), volume 6(9); http://chango.ibmb.csic.es/ARCIMBOLDO
6. Panjikar S. *et al.*, *On the combination of molecular replacement and single anomalous diffraction phasing for automated structure determination*, Acta Crystallogr (2009), **D**65; http://www.embl-hamburg.de/Auto-Rickshaw
7. *The CCP4 Suite: Programs for Protein Crystallography*, Acta Crystallogr. (1994), **D**50
8. McCoy, A.J. *et al.*, *Phaser crystallographic software*, J. App. Cryst. (2007), vol. 40