High throughput phasing with SHELXC/D/E

SHELXC is designed to provide a simple and fast way of setting up the files for the programs SHELXD (heavy atom location) and SHELXE (phasing and density modification) for macromolecular phasing by the MAD, SAD, SIR and SIRAS methods. These three programs may be run in batch mode or called from a GUI such as *hkl2map* (Pape & Schneider, *J. Appl. Cryst.* **37** (2004) 843-844; available from Thomas R. Schneider: Schneider@ifom-firc.it). SHELXC is much less versatile than the Bruker XPREP program for this purpose, but if you are sure of the space group and there are no problems with the indexing or twinning and the *f'* and *f"* parts of the scattering factors do not need to be refined, SHELXC should be adequate. SHELXC can read either HKL2000 .sca files or SHELX .hkl files (*F*-squared unless the -f switch is used to specify *F*). To transfer data from CCP4 it is advisable to generate .sca files using 'output unmerged polish' from SCALA or to use the program mtz2sca written by Tim Grüne and supplied with SHELX. The current version of SHELXC outputs extra useful diagnostic statistics if fed *unmerged* data. SHELXC, SHELXD and SHELXE are stand-alone executables that do not require environment variables or parameter files etc., so all that is needed to install them is to put them in a directory that is in the 'path' (e.g. /usr/local/bin or ~/bin under Linux).

SHELXC reads a filename stem on the command line plus some instructions from 'standard input'. It writes some statistics to 'standard output' and prepares the three files needed to run SHELXD and SHELXE. It can be called from a GUI using a single command line such as:

```
shelxc xx <t
```

which would read the instructions from the file t and write the files xx.hkl (h,k,l,I,$\sigma(I)$ in SHELX HKLF4 format for density modification by SHELXE), xx_fa.ins (cell, symmetry etc. for heavy atom location using SHELXD) and xx_fa.hkl (h,k,l,F_A,$\sigma(F_A)$,α for both SHELXD and SHELXE). The starting phases for density modification are estimated as (heavy atom phase + α) in the simplified approach used by SHELXE, α is calculated by SHELXC from the anomalous and dispersive differences. For SAD α is 90º ($I_+ > I_-$) or 270º ($I_+ < I_-$), for SIR and RIP α is 0º or 180º and for SIRAS or MAD α may be anywhere in the range 0º to 360º.

The above command line could be used under UNIX or Windows; under UNIX the commands to run SHELXC, SHELXD and SHELXE and the instructions for SHELXC may also be combined into a single script file as shown in the following examples. In these scripts, the instructions start on the line after '<<EOF' and are terminated by 'EOF'. The instructions may be given in any order; CELL (unit-cell), SPAG (space group in PDB notation, spaces are ignored) and FIND (followed by the number of heavy atoms) must be given; the optional instructions SFAC, MIND, NTRY, SHEL, ESEL and DSUL, if present, are copied to the SHELXD input file.

MAD example

```
shelxc jia <<EOF
NAT jia_nat.hkl
HREM jia_hrem.sca
PEAK jia_peak.sca
INFL jia_infl.sca
LREM jia_lrem.sca
CELL 96.00 120.00 166.13 90 90 90
SPAG C2221
FIND 8
NTRY 10
EOF
shelxd jia_fa
shelxe jia jia_fa -s0.6 -m20
shelxe jia jia_fa -s0.6 -m20 -i
```

In this example (kindly donated by Zbigniew Dauter; Li et al., *Nature Struct. Biol.* **7** (2000) 555-559), Se-Met MAD data at four wavelengths are used to calculated the F_A-values and phase shifts α that are written to the file jia_fa.hkl. The native (S-Met) data are read from jia_nat.hkl and written to jia.hkl. The file jia_fa.ins is prepared using the given cell, space group, FIND and NTRY instructions as well as a suitable SHEL command to truncate the resolution. SHELXD then searches for 8 (FIND) selenium atoms using 10 attempts (NTRY), and SHELXE is run for 20 cycles (-m) of density modification for both heavy atom enantiomorphs (-i inverts) with a solvent content (-s) of 0.6. The protein phases are written to jia.phs and jia_i.phs resp. If NAT is not specified, SHELXC would analyze the four MAD datasets to generate the (SeMet) native data jia.hkl, in which case -h should be specified for SHELXE since the selenium atoms are present in the 'native' structure. For MAD at least two wavelengths are required, at least one of which should be PEAK or INFL.

SAD Example

This example of thaumatin phasing by means of the native sulfur anomalous signal (Debreczeni et al., *Acta Cryst.* **D59** (2003) 688-696) uses 1.55 Å in-house Cu$K\alpha$ data:

```
shelxc thau <<EOF
SAD thau-nat.hkl
CELL 58.036 58.036 151.29 90 90 90
SPAG P41212
FIND 9
DSUL 8
MIND –3.5
NTRY 100
EOF
shelxd thau_fa
shelxe thau thau_fa -h -s0.5 -m20
shelxe thau thau_fa -h -s0.5 -m20 –i
```

The anomalous differences are extracted from the native data so only one data file is required. The sites specified by FIND consist of one methionine and 8 super-sulfurs, which are then resolved into disulfides using the DSUL instruction that is passed on to SHELXD (Debreczeni *et al., Acta Cryst.* **D59** (2003) 2125-2132). Alternatively one could try to find the individual sulfurs with:

```
SHEL 999 2.0
FIND 17
MIND –1.7
```

Here the resolution cutoff has been reduced from 2.1 Å (which SHELXC would have suggested) to 2.0 Å to improve the chances of resolving the sulfurs. The SHEL, FIND, MIND and NTRY instructions are transferred to the file thau_fa.ins for the sulfur atom location with SHELXD. Note that the phases can be improved further in this case by using more SHELXE cycles than the usual 20.

SIRAS example

This involves the solution of the thaumatin structure using the above 1.55 Å data as native and 2.0 Å Cu*K*α data from a quick iodide soak. SIRAS usually gives the best results for iodide soaks, but it is also possible in this case to use SIR (change 'SIRA' to 'SIR') or iodine SAD (change 'SIRA' to 'SAD').

```
shelxc thaui <<EOF
NAT thau-nat.hkl
SIRA thau-iod.hkl
CELL 58.036 58.036 151.29 90 90 90
SPAG P41212
FIND 17
NTRY 10
MIND –3.5 –0.1
EOF
shelxd thaui_fa
shelxe thaui thaui_fa -s0.5 –m20
shelxe thaui thaui_fa -s0.5 –m20 -i
```

Critical parameters for SHELXD

In general the critical parameters for locating heavy atoms with SHELXD are:

1. The resolution cutoff. In the MAD case this is best determined by finding where the correlation coefficient between the signed anomalous differences for wavelengths with the highest anomalous signal (PEAK and HREM or PEAK and INFL) falls below about 30%. For SAD a less reliable guide is where $\Delta F/\sigma F$ falls below about 1.2 (a value of 0.8 would indicate pure noise), and for S-SAD with Cu*K*α the data can be truncated where I/σ for the native data falls below 30. If unmerged data are used, SHELXC calculates a correlation coefficient between two randomly selected subsets of the signed

anomalous differences; this is a better indicator because it does not require that the intensity esds are on an absolute scale, but it does require a reasonable redundancy and again the data can be truncated where it drops to below 30% (the CCP4 program SCALA prints a similar statistic).

2. The estimated number of sites (FIND) should be within about 20% of the true number. For SeMet or S-SAD phasing there should be a sharp drop in the occupancy after the last true site. For iodide soaks, a good rule of thumb is to start with a number of iodide sites equal to the number of amino-acids in the asymmetric unit divided by 15. If after SHELXD occupancy refinement the occupancy of the last site is more than 0.2 it might be worth increasing this number, and *vice versa*.

3. A common 'user error' is to set MIND -3.5 even though the distances between heavy atoms are less than 3.5 Å. For example, in a Fe_4S_4 cluster the Fe...Fe distance is about 2.7 Å, so MIND -2 would be appropriate. A disulfide bond has a length of 2.03 Å so then MIND -1.5 could be used to resolve the sulfur atoms, however if DSUL is used for this purpose MIND -3.5 is required.

4. If heavy atoms can lie on special positions (as is the case with an iodide soak in a space group with twofold axes) the rejection of atoms on special positions should be switched off by giving the second MIND parameter as -0.1 (as in the above thaumatin example).

5. In cubic space groups the Patterson seeding (PATS) is slow and less effective, it is recommended that 'PATS' is replaced by 'WEED 0.3'.

For MAD, a CC of 40 to 50% indicates a good solution, for SAD etc. values around 30% may well be correct, especially if the same solution or group of solutions has the highest values of CC, CC(Weak) and PATFOM, and they are well separated from the values for the non-solutions. The CC values tend to increase as the resolution is lowered. Heavy atom soaks truncated to low resolution often give spuriously high CC values, ***but these 'solutions' can be recognized as false by their low CC(weak) values***.

In difficult cases SHELXD can be run with different SHEL instructions, e.g. truncating the data in steps of 0.1 Å, and the CC values compared. This is especially convenient if a computer farm can be used to run the jobs in parallel. If the best CC is plotted against the resolution, a local maximum (when also observed for the CC(weak) values) may indicate a correct solution.

The default weights for the CC are $1/\sigma(E)^2$. The presence of one or two reflections with very low esds can lead to unreasonably high values of the CC for wrong solutions. If the esds are unreliable it is advisable to use 'CCWT 0.1' in the .ins file for SHELXD. The precision of the heavy atom coordinates can be improved, at the cost of the CPU time, by making the Fourier grid finer (e.g. FRES 4 instead of the default 2.5).

Modes of operation of SHELXE

SHELXE has following modes of action (xx and yy are filename stems):

`shelxe xx` [reads xx.hkl and xx.ins, phases from atoms]
`shelxe xx yy` [**normal mode**: reads xx.hkl, yy.hkl, yy.res]
`shelxe xx.phi` [reads xx.phi, xx.hkl, xx.ins]
`shelxe xx.fcf` [reads only xx.fcf]
`shelxe xx.phi yy` [reads xx.phi, xx.hkl, xx.ins, yy.hkl]
`shelxe xx.fcf yy` [reads xx.fcf, yy.hkl, yy.res]

xx.hkl contains native data, yy.hkl contains F_A and α and should have been created using SHELXC or XPREP. xx.phi has .phs format (*h*,*k*,*l*,*F*,*fom*,ϕ in free format) and can be made by renaming a .phs output file from SHELXE, but only the starting phases are read from it; if a .phi file is read, the cell and symmetry are read from xx.ins and the native F-values are read from xx.hkl. xx.fcf (from a SHELXL structure refinement) provides cell, symmetry and starting phases. The output phases are written to xx.phs, the log file is written to xx.lst and, if -b is set, improved substructure phases are output to xx.pha and revised heavy atoms to xx.hat.

The first two modes provide density modification starting from atoms or phases, the third and fourth modes are for phase extension, the fifth is an inverse cross-Fourier for finding heavy atoms for a second derivative (yy) with the same origin as the first (xx), and the last mode is useful to confirm the heavy atom substructure from the final refined phases. This is useful as a post-mortem if SAD or MAD phasing fails but the structure could be solved by other means. For these last two modes, the phases for the inverse Fourier are ($\phi_{nat} - \alpha$), where ϕ_{nat} may be refined (-m etc.) and α is taken from yy.hkl. A few cycles of phase refinement may reduce the noise in such maps by improving the weights.

Phasing and density modification with SHELXE (the normal mode)

SHELXE normally requires a few command line switches, e.g.

`shelxe xx yy -m20 -s0.45 -h8 -b`

would do 20 cycles density modification with a solvent content of 0.45, phasing from the first 8 heavy atoms in the yy.res file from SHELXD assuming that they are also present in the native structure (-h8), and then use the modified density to generate improved heavy atoms (-b). The switch -i may be added to invert the substructure (and if necessary the space group), this writes xx_i.phs instead of xx.phs etc., and so may be run in parallel.

A big difference in the *contrast* between the two heavy-atom enantiomorphs usually indicates a good SHELXE solution. However in the case of SIR, both have the same contrast but one gives the inverted protein structure. The contrast is also the same

for both if the heavy-atom substructure is centrosymmetric. In the case of SAD both heavy atom enantiomers then give the correct structure, for SIR the result is an uninterpretable double image.

The pseudo-free correlation coefficient (based on the comparison of E_o and E_c for 10% of the data left out at random in the calculation of a map that is then density modified and Fourier back-transformed in the usual way) is now printed out before every *N*th cycle (set by -j, the default is -j5); a value above 70% usually indicates an interpretable map. The pseudo-free CC (which is also reported in the hkl2map plot of contrast against cycle number) is also a good indication as to when the phase refinement has converged.

The solvent content (-s) is by far the most critical parameter for SHELXE, and it is often worth varying it in steps of about 0.05 to maximize the difference in contrast between the two enantiomorphs and the 'pseudo-free CC' (another application for a computer farm!). Usually the optimal solvent content is higher than the calculated value at low resolution (disordered side-chains?) and lower at high resolution (ordered solvent?). Sometimes it is necessary to use many (several hundred) cycles (-m) if the starting phase information is weak but the resolution is very high. For low resolution data, the use of more than 20 refinement cycles is normally counter-productive. The current values of all parameters are output at the start of the SHELXE output, the default values of other parameters will rarely need changing.

The -b switch in SHELXE causes updated heavy atom positions to be written to the file name.hat (or name_i.hat). This file can be copied or renamed to the .res file (which should be saved first!) and used to recycle the heavy atoms. Versions 0.0.34 and later of the graphics program Coot (Emsley & Cowtan, *Acta Cryst.* **D60** (2004) 2126-2132) should be able to deduce the space group name from the symmetry operators in this file, and so a very convenient way to obtain a map after running SHELXE is to start Coot, read in 'coordinates' from the .hat or _i.hat file, and then input the phases from the .phs or _i.phs files and the phases of the heavy atom substructure from the .pha or _i.pha files. It is normally necessary to increase the sigma level of the latter map (by hitting '+' several times). This procedure even works correctly when the space group has been inverted by SHELXE!

Good quality MAD data, a high solvent content and/or high resolution for the native data can lead to maps of high quality that can be autotraced (e.g. with wARP) immediately. The .phs files contain *h*, *k*, *l*, *F*, *fom*, ϕ and $\sigma(F)$ in free format and can be read directly into Coot or converted to CCP4 .mtz format using f2mtz, e.g. for further density modification exploiting NCS using the CCP4 program Pirate. Note that if the inverted heavy atom enantiomorph is the correct one, the corresponding phases are in the *_i.phs file and SHELXE may have inverted the space group (e.g. $P4_1$ to $P4_3$), which should be taken into account when moving to other programs!

The free lunch algorithm (FLA)

The new switch -e may be used to extrapolate the data to the specified resolution (the 'free lunch algorithm', based closely on work by the Bari group: Caliandro *et al*., *Acta Cryst*. **D61** (2005) 556-565); -e1.0 can produce spectacular results when applied to data collected to 1.6 to 2.0 Å, but since a large number of cycles is required (-m400) and the 'contrast' and 'connectivity' become unreliable (the pseudo-free CC is the only reliable map quality indicator when the FLA is used), it may be best to establish the substructure enantiomorph and solvent content without -e first. The default setting when -e is not specified is to fill in missing low and medium resolution data but not to extrapolate to higher resolution than actually measured (to switch off this filling in, use -e999). The resolution requirements for the FLA still need to be explored, but so far there have been no reports of it causing a deterioration in map quality, and in a few cases the mean phase error was reduced by as much as 30º relative to density modification without it.

RIP with SHELXC/D/E

RIP (radiation damage induced phasing) can be regarded as a sort of isomorphous replacement where the 'after' dataset has lost a few atoms that are particularly susceptible to radiation damage. In fact, many structures have solved unintentionally with a helping hand from RIP! In a MAD experiment, ***provided that the 'inflection point' dataset is collected last from the same crystal***, the radiation damage has the effect of making *f'* for the MAD element at this wavelength even more negative than usual, enhancing the dispersive part of the MAD signal. This is especially true of bromine MAD on bromouracil derivatives, because the radiation near the bromine absorption edge appears to be particularly effective at breaking the bromine-carbon bonds irreversibly. Of course if the inflection data are collected first the RIP and dispersive component of the MAD signal will tend to cancel one another, causing the MAD analysis to fail, although SAD may still be able to solve the structure (also a common scenario).

RIP (without using anomalous scattering) or RIPAS (like SIRAS, assuming that the anomalous atoms are also those most sensitive to radiation damage) can be capable of solving difficult structures. A typical procedure on a third generation synchrotron beamline is to collect the 'before' dataset with an attenuator in the beam, then to fry the crystal for a couple of minutes with the unattenuated beam, and finally to collect an 'after' dataset with the attenuator in. In the SHELXC instructions, the 'before' data are called 'NAT' or 'BEFORE' and the 'after' data are called 'RIP' or 'AFTER'. The critical parameter is the scale factor applied to the 'after' data after both datasets have been brought onto a common scale. This is set by the SHELXC instruction 'DSCA' and should usually be in the range 0.95 to 1.00. This scale factor may also be used for SIR and SIRAS, where it is applied to the native data, but it appears to be less critical than for RIP. For RIPAS, the 'after' data should be called 'RIPA' and the 'RIPW' instruction specifies the weight *w* (default 0.6) for the anomalous contribution from the 'before' dataset (a weight 1–*w* is applied to the 'after' data).

In RIP or RIPAS phase determination is usually necessary to recycle the 'heavy atom' sites by renaming the output .hat (or _i.hat) file as .res and rerunning SHELXE. It is advisable to edit this file so as to retain the stronger negative sites, these may well correspond to the new positions of displaced atoms. SHELXE can read negative occupancies but SHELXD can only search for positive atoms. It should be noted that in a pure RIP experiment, both hands of the radiation damage substructure will give the same figures of merit, but one will lead to an electron density map that is a mirror image of the true map (the helices will go the wrong way round). Extensive details of the use of SHELXC, SHELXD and SHELXE for RIP phasing may be found in Nanao, Sheldrick & Ravelli, *Acta Cryst.* **D61** (2005) 1227-1237.

Obtaining the SHELX programs

SHELXC/D/E and test data may be downloaded from the SHELX fileserver. The application form should be printed out from http://shelx.uni-ac.gwdg.de/SHELX/ This form should be completed and faxed to +49-551-392582. Downloading instructions will then be emailed to the address given on the form. The programs are free to academics but a small license fee is required for 'for-profit' use. This fee of $2499 (US) or 1999 Euros (rest of the world) covers the use of the current versions (plus future minor updates) on an unlimited number of computers for an unlimited time at one geographical location. It covers the costs of supporting all users, we do not make a profit.

References

If these programs prove useful, you may wish to cite:

Sheldrick, G.M., Hauptman, H.A., Weeks, C.M., Miller, R. & Usón, I. (2001). "*Ab initio* phasing". In *International Tables for Crystallography, Vol. F,* Eds. Rossmann, M.G. & Arnold, E., IUCr and Kluwer Academic Publishers, Dordrecht pp. 333-351. [*Full background to the dual-space recycling used in SHELXD*].

Schneider, T.R. & Sheldrick, G.M. (2002). "Substructure Solution with SHELXD", *Acta Crystallogr.* D**58** 1772-1779. [*Heavy atom location with SHELXD*].

Sheldrick, G.M. (2002), "Macromolecular phasing with SHELXE", *Z. Kristallogr.* **217**, 644-650. [*The definitive reference for SHELXE, usually cited wrongly*].

I am able to supply pdfs of the last two of these on request.

George M. Sheldrick, December 30th 2005.