SHELXD and SHELXE

The structure solution program SHELXD (called XM in the Bruker SHELXTL system, but identical to SHELXD except in the logo) is able to solve larger *ab initio* problems than SHELXS-97, and is also useful for locating the heavy atoms or anomalous scatterers from SIR, SAD, SIRAS or MAD data. From January 2002 SHELXD is available as source and precompiled binaries for common operating system as part of the SHELX-97 system. XM is available from Bruker Nonius as part of the SHELXTL system, which includes the whole of SHELX plus programs not in the public domain such as the interactive molecular graphics program XP and reflection data manipulation program XPREP. In this documentation both XM and SHELXD will be referred to as SHELXD.

For the MAD, SAD, SIR etc. applications of SHELXD the location of the heavy atom sites is only one step in the structure solution. The new program SHELXE (XE in the Bruker SHELXTL) can read the *.res* file containing the heavy atom sites written by SHELXD and estimate the native phases and the corresponding weights (figures of merit). SHELXE outputs the phases in an XtalView format *.phs* file so that a map can be viewed using interactive graphics or the phases can be improved by density modification using program such as DM, SOLOMON, RESOLVE etc. SHELXE is robust, fast and simple to use, but it must be emphasized that the resulting phases may be inferior to those produced by much more sophisticated maximum likelihood programs such as SHARP, SOLVE or MLPHARE. However in favorable cases it may even prove possible to autotrace the maps from SHELXE directly, e.g. using wARP.

For SIR and SAD problems SHELXE starts with the centroid phases from the Harker construction (Harker, 1956); for MAD and SIRAS an unambiguous phase can be assigned. Sigma-A weights (Read, 1986) are used throughout. In the case of SAD and SIR a single density truncation cycle retaining only about 7% of the density is applied to resolve the twofold ambiguity for appropriate reflections; this is similar to the *low density elimination* used by Woolfson et al. ( ) and to a density modification procedure proposed by Giacovazzo & Siliqi (1997).

The crude density modification performed by SHELXE may be termed the *sphere of influence* method. A sphere of radius 2.42Å (a typical 1,3-distance in virtually all organic and macromolecular structures) is constructed around each pixel of the map, and the variance of the electron density around a given pixel is calculated using 92 (or 272) optimally distributed pixels that lie close to this sphere. The variances are sorted but instead of using them to define a sharp solvent boundary, a fuzzy boundary is generated so that pixels with very high sphere variances will be entirely in the 'protein' region and those with very low variances will

be entirely in the 'solvent' region and the ones in between are assigned probabilities between 0 and 100% that they are in the solvent region. In the protein region the negative density is reset to zero and in the solvent region it is 'flipped' (Abrahams & Leslie, 1996). A pixel that has been assigned a 60% probability of being in the solvent region is assigned a 60:40 weighted average of the densities resulting from the solvent and protein treatments.

It was anticipated that by using a little chemical knowledge (the 1,3-distance) it would be possible to improve maps given very high resolution data, but in practice the method still works well with 3Å data provided that the solvent content is relatively high. For very high resolution data (better than 1.5Å) or very high solvent content (>60%) the SHELXE phases can have rather high map correlation coefficients (>0.9) with the phases from the final refinement. In less favorable cases it may well be possible to improve the phases further using other more sophisticated density modification programs, especially if non-crystallographic symmetry (NCS) can be exploited. An attempt is made to estimate realistic weights (foms) in SHELXE so that further phase refinement using other programs is facilitated.

SHELXE is currently a beta-test that is being made available in precompiled form without extra license fees etc. but with an expiry date (1/1/03) to registered SHELX and SHELXTL users. If it proves successful it will be incorporated in future SHELX and SHELXTL releases that will have to be licensed separately.

Introduction to SHELXD

SHELXD is a stand-alone executable and does not require any other program, initialization files or environment variables etc. The input to SHELXD consists of two files, *name.ins* and *name.hkl*, both of which can conveniently be created using the Bruker Nonius XPREP program. The *.hkl* file has the standard SHELX format and with the exception of two or three instructions in the *.ins* file is very similar to the input for SHELXS. SHELXD expects ONE and only one source of starting atoms. This can take the form:

A: Input atoms in normal SHELX format for expansion using **PLOP**

B: PATS for Patterson seeding of the dual-space direct methods

C: GROP and a PDB-format model for fragment seeding

D: Random atoms (used if none of the above apply)

For substructure solution using MAD data etc. option **B** (**PATS** + **FIND** but no **PLOP**) is recommended. In each case the action is specified in the *.ins* file that also contains crystal data in the usual SHELX form. The reflection data consists of an *.hkl* file containing F^2 (**HKLF 4**) or *F*-values (**HKLF 3**). These may correspond to either native data for *ab initio* structure solution or structure expansion, or MAD, SAD, SIR or SIRAS F_A or ΔF values for heavy or anomalous atom location.

Dual-space recycling (Miller et al., 1993; Miller et al., 1994; Sheldrick et al., 2001), using the largest *E*-values (**FIND**) is followed by *peaklist optimization* (**PLOP**; Sheldrick & Gould, 1995); one or both of these commands must be present. In the case of structure expansion only **PLOP** can be used and the program then stops. When the starting atoms are generated randomly or by **PATS** or **GROP**, the calculations are repeated with new sets of starting atoms each time. The total number of such tries may be specified with **NTRY**, otherwise the program runs for ever (unless interrupted by a *name.fin* file).

When the final correlation coefficient CC (after **PLOP**) for an atomic resolution *ab initio* run of SHELXD is 65% or greater, the structure is almost certainly solved. SHELXD writes the best solution so far to a SHELX format file *name.res* and a PDB format file *name.pdb*. The former can be examined with the interactive graphics program XP that is part of the Bruker SHELXTL system. If XP is not available the PDB file may be displayed with RASMOL (use the ball and stick display mode). Note that this may be done before stopping SHELXD. If the structure is clearly solved, SHELXD may be terminated cleanly by creating a file *name.fin* in the working directory.

Examples of *ab initio* structure solution with SHELXD

To illustrate full structure solution by *ab initio* methods, a test example is provided (in the *egs* subdirectory on the SHELX ftp site) in the form of the files *pn1a.ins* and *pn1a.hkl*. Four different ways of solving the structure are included in the *.ins* file; in order to run the various tests it will be necessary to comment out some lines (by putting a space character at the beginning of the line). The file is read only as far as the first **HKLF** instruction. This test structure was kindly provided by Jenny Martin, University of Queensland, Australia. It consists of (GCCSLPPCAANNPDYC), a linear polypeptide with two disulfide bridges, giving 110 non-hydrogen peptide atoms plus 12 solvent atoms. The space group is $P2_1$ and the resolution of the data 1.1Å. For further details see Hu et al. (1996). In the following examples, **TITL...UNIT** in the normal SHELX format is assumed at the start of the *.ins* file and **HKLF 4** (or **HKLF 3**) followed by **END** at the end of the file. The cell contents defined by **SFAC** and **UNIT** are only used by **PLOP**; in the **FIND** stage the atoms are assumed to be of the same type but with occupancies proportional to the square root of the peak height, unless occupancy refinement is used (**TANG** with a negative first parameter).

```
FIND 80
PLOP 120 140 160
NTRY 50
```

This will search (**FIND**) for 80 atoms in the dual-space stage; it is usually more efficient to search for ca. 25% less than the total number of non-solvent atoms, especially when - as here - some heavier atoms such as sulfur are present. In the **PLOP** stage on the other hand one should specify more than the expected number of atoms because this procedure involves the elimination of the 'wrong' atoms. One can leave **NTRY** out in which case the job will run forever (unless aborted or stopped more gently by creating a *name.fin* file in the same directory).

An alternative approach is to use Patterson seeding instead of random starting atoms. One can then look for say 80 atoms as above with **FIND**, or alternatively first optimize the sulfur substructure (in this case four atoms) with **FIND** and expand to the full structure with **PLOP**. The Patterson seeding may be performed for example with a randomly oriented fixed length vector (for a disulfide bond). Everything after a '!' sign in a SHELX *.ins* file is treated as a comment.

```
PATS -2.06  ! S-S distance
PSMF -4     ! supersharp Patterson
FIND 4 5
MIND -1.8   ! S-S > 1.8A, calc. PATFOM
TEST 10 5
PLOP 50 80 120 160 160
NTRY 20
```

Alternatively the Patterson seeding may use the highest Patterson peaks as translation search vectors:

```
PATS
PSMF -4
FIND 4 5
MIND -1.8
TEST 10 5
PLOP 50 80 120 160 160
NTRY 20
```

Patterson or fragment seeding does not have to go through the **FIND** stage to optimize the atomic positions, though this is strongly recommended and has the advantage that all four sulfurs can be used. It is also possible to go into structure expansion with **PLOP** directly, and this facility can be tested using the two-atom disulfide fragment as follows. It should be noted that two sulfur atoms are quite adequate for **PLOP** to expand to the full structure, but the CC threshold (the first **TEST** parameter) for entering the **PLOP** stage needs to be reduced a little (in the above tests, it had the default of 45% for **FIND 80** and was set to 10 for **FIND 4**).

```
GROP
TEST 8 5
PLOP 30 50 80 120 160 160
NTRY 20
ATOM      1  S   CYS     1       0.000   0.000   0.000 1.000 10.00
ATOM      2  S   CYS     1       0.000   0.000   2.060 1.000 10.00
```

The two sulfur atoms are given in fixed PDB fixed format. As a further example (not provided as test files) of seeding based on an initial fragment search, for a cyclodextrin structure with four beta-cyclodextrins in the asymmetric unit and with data barely to atomic resolution, the following could be tried:

```
GROP
FIND 240
PLOP 320 400
ATOM      1  C41 MOL     1      -3.859   4.863   7.904 1.000 10.00
ATOM      2  C31 MOL     1      -5.081   4.209   8.524 1.000 10.00
ATOM      3  C21 MOL     1      -5.211   2.740   8.155 1.000 10.00

   ... diglucose fragment in PDB format ....

ATOM     21  C52 MOL     1      -0.292   4.714   7.025 1.000 10.00
ATOM     22  O52 MOL     1      -0.642   5.837   6.253 1.000 10.00
```

A major new facility in SHELXD for small molecules is the ability to solve merohedrally twinned structures by *ab initio* methods; all that is required is to input the SHELXL instructions **TWIN** and estimated **BASF** parameter (which is held at a fixed value throughout). XPREP can be used to find the **TWIN** matrix and estimate the **BASF** parameter value. **TWIN** and **BASF** are only applied at the **PLOP** stage, and are ignored by **PATS**, **GROP** and **FIND**.

Macromolecular phasing using SHELXD and SHELXE

SHELXE is intended to be run immediately after SHELXD. It picks up the *.res* file containing the best substructure solution (so far) from SHELXD. Since very few parameters are required for SHELXE they are all given on the command line. When the correlation coefficients indicate that SHELXD has 'solved' the substructure, it can be terminated (by writing a dummy *name.fin* file into the working directory - under UNIX the **touch** instruction can be used for this) and TWO SHELXE jobs started. Two jobs are almost always necessary because the heavy atom substructure and where appropriate the space group may have to be inverted; there is a 50% chance that the heavy atom enantiomorph will be wrong! The command lines for these two jobs are identical except that one contains the **-i** switch. These two jobs may be run simultaneously because the files do not clash; the **-i** job adds '_i' to the end of the first part of the filename for the output files. Often it will become clear from the console output which heavy atom enantiomorph is correct (see examples below) and the other job can be killed with <ctrl-C>.

Before phasing with SHELXD and SHELXE it is necessary to prepare three input files: *name-df.ins*, *name-df.hkl* and *name.hkl*. The first two are read by SHELXD, the last two by SHELXE, which also reads the file *name-df.res* written by SHELXD. Up to the period, the filename can be freely chosen but must be the same for the first two files; see the examples below. All three files can conveniently be set up using the Bruker XPREP program, but the information below should enable other sources to be used. Note that Bruker Nonius are often willing to provide a free demo version of XPREP (fully featured but with an expiry date), anyone interested should contact sbyram@bruker-axs.com, trixie.wagner@bruker-axs.de or anita.coetzee@nonius.nl.

The *name-df.ins* file contains (at least) the following instructions in the order given:

TITL (followed by any title on the same line)
CELL l **a b c** a b g (in Å and deg.: λ is ignored but is standard for SHELX)
LATT and **SYMM** (to define the space group, see examples and the SHELX manual)
SFAC Se (or any other single element, even if there are several heavy atom types)
UNIT M (approximate number of heavy atoms per cell multiplied by 4)
SHEL 999 d (where **d** is the resolution at which to truncate the data)
PATS (Patterson seeding)
FIND N (number of sites to search for, should be within 20% for best results)
MIND -3.5 (minimum allowed distance between sites)
HKLF 3 (to read F rather than F^2)
END

The critical parameters are **d**, the resolution at which to truncate the data, and **N**, the number of atoms to be searched for; it may be worth trying different values of these two parameters

in difficult cases.

The optimal value of **d** may be estimated using XPREP, either from the mean ratio of ΔF to its esd (assuming that the data have been processed so that the esds are on an absolute scale, i.e. χ^2 is close to one), or from the correlation coefficient between the signed anomalous differences for two datasets (different MAD wavelengths or in the case of SAD different crystals). It should be noted that there is almost always an optimal value of **d** and it should be larger than the resolution limit of the diffraction pattern. Often 3Å to 3.5Å gives good results for MAD phasing. If XPREP is not available then a good rule of thumb is to set **d** to 0.5Å less than the diffraction limit.

At the end of the dual-space direct methods SHELXD refines the site occupancies assuming that all atoms are of the same type. This provides an adequate approximation in the case where different anomalous scatterers are present (e.g. Ca^{2+} and S in the trypsin example discussed below). It also shows when the actual number of sites is different from the value input on the **FIND** instruction; for a selenomethionine MAD experiment there should be a clear drop in occupancy after the last site. For halide soaks on the other hand there is often a continuous descent to the noise level reflecting the variable occupancies of the sites. The occupancy refinement is switched on by a negative first **TANG** parameter; this is the default if there is no **PLOP** instruction.

The cell contents (**SFAC/UNIT**) should be specified correctly when SHELXD is used for full *ab initio* structure solution, but for substructures a single element type should be specified and the number of sites expected per cell multiplied by about four so that the probabilities are calculated correctly for the *minimal function* and *Ralpha* figures of merit. Since these are only printed as information - the correlation coefficient alone is used to decide which solution is 'best' - the **SFAC/UNIT** parameters are not important for substructure solution.

For large selenomethionine substructures (which behave more like equal atom *ab initio* structure solution of small molecules) it may be worth increasing the number of Patterson peaks used for the Patterson seeding (e.g. **PATS 200**; the default is 100) and adding the instructions **WEED 0.3** (random omit maps) and **SKIP 0.5** (uranium atom removal). The latter two are the defaults when **PLOP** is present but are switched off by default if **PLOP** is absent. When **PATS** is used, **WEED** produces a much smaller additional improvement in the hit ratio than when **PATS** is absent. For small substructures (<10 sites), **WEED** and **SKIP** can do more harm than good by eliminating too many correct sites at once.

The minus sign for the first **MIND** parameter specifies that the PATFOM figure of merit and crossword table should be calculated. For phasing using the anomalous scattering of sulfur, a distance of about 1.7Å is required if the resolution of the ΔF data (as truncated using **SHEL**)

permits the sulfur atoms in disulfide bridges to be resolved from each other (see trypsin example below). The default option in the **FIND** stage of SHELXD is to ignore all sites on special positions; to include possible sites on special positions, set the second **MIND** parameter to -0.1. This can happen for halide soaks etc. but is not required for the two examples below (selenomethionine cannot lie on a special position, and there are no special positions in $P2_12_12_1$).

It may also be worth adding **NTPR 100** or **NTPR 1000**, otherwise the SHELXD job will never finish. Alternatively **NTPR** can be left out and the job terminated by creating a *name-df.fin* file.

The file *name-df.hkl* consists of one line per reflection, terminated by the end of the file or by a line with all numbers zero. It is read using the FORTRAN format 3I4,2F8.2,I4; as normal when reading floating point numbers with FORTRAN, the number of figures after the decimal point may be varied but the numbers must be contained within the 8 character fields and the decimal point must be present in the number. Each line consists of *h*, *k*, *l*, [ΔF or F_A], [$\sigma(\Delta F)$ or $\sigma(F_A)$] and α, where α is the estimated phase shift in degrees that has to be added to the heavy atom phase to give the native protein phase. ΔF or F_A are always given as positive numbers. In the SIR case, α is zero if the derivative F is greater than the native F and 180 if the opposite is true; for SAD, α is 90 if $F_+ > F_-$ and 270 if $F_+ < F_-$ For MAD or SIRAS data, α may be anywhere in the range 0 to 360. α is only read by SHELXE, not by SHELXD.

The file *name.hkl* contains *h*, *k*, *l*, F^2 and $\sigma(F^2)$ in format 3I4,2F8.2 for the native data and is terminated by the end of the file or by a line with all numbers zero. In a selenomethionine MAD experiment it could either be a remote wavelength or (as in the example below) it could be the data from the native (methionine) crystal if that diffracted to higher resolution. Usually the same data will be used for the final refinement of the structure.

After starting SHELXD (with the command line **shelxd name-df**) the program first prints a summary of all parameters used, then calculates and stores the Patterson and the phase relations for the tangent formula. The solution with the best *CC* (correlation coefficient) so far is written to the *name-df.res* file. One should wait until there are one or more solutions with *CC* and *CC(weak)* at least 30 and 15 resp. and well separated from the rest, but in practice it is worth waiting a few minutes longer in case there is an even better solution. When it appears (from the *CC* values and the other figures of merit) that the substructure has almost certainly been 'solved' , SHELXD can be terminated or interrupted and the two SHELXE jobs started. Although SHELXD can in principle be left running throughout, it is better to force at least a pause so that both SHELXE jobs pick up the same *name-df.res* file. Otherwise if a new marginally better solution is written to the *name-df.res* file in the meantime it could happen that both SHELXE jobs correspond to the same heavy-

atom enantiomorph instead of opposite ones! SHELXE writes the phases to *name.phs* (ready for input to XtalView or - via a CCP4 script - DM) and a listing file *name.lst*. All SHELXE parameters are specified on the command line, e.g.

```
shelxe name name-df -h -m10 -s0.55
```

could be used to pick up the cell, symmetry and atoms from *name-df.res* written by SHELXD (there is really no need to look at the contents of this file !!).

The **-h** flag tells the program that the 'heavy atoms' are also present in the native data. This will be true for a selenomethionine MAD experiment when the remote wavelength is also used as native, and is always true for anomalous sulfur phasing. For SIR on say an iodide soak **-h** would not be used because there are no iodides in the native structure.

The **-m** flag gives the number of density modification cycles. This should be sufficient for convergence of the *contrast* and *connectivity* parameters printed out by SHELXE: **-m10** is a good first guess.

The **-s** flag gives the solvent content (default is 0.45). The solvent content is the most critical parameter and should be estimated carefully; a good way is to assume that an average amino-acid has a volume of 140Å^3. If it is difficult to estimate the solvent content it may be worth running the program with several different values. Note that there should be no spaces between flags and numerical values.

A **-i** flag should be used in the second SHELXE run to force inversion of the heavy atoms and (if required) space group (remember this when running XtalView etc.!). This writes the files *name_i.phs* and *name_i.lst* and so does not clash with the first SHELXE run. Thus the second run may be started with:

```
shelxe name name-df -h -m10 -s0.55 -i
```

The values of the *contrast* and *connectivity* printed out by SHELXE at the end of each density modification cycle should be compared. The *contrast* is high when some regions of the map show large fluctuations in density and others are fairly flat; the value calculated is related to the standard deviation of the local r.m.s. electron density as proposed by Terwilliger & Berendzen (1999). The *connectivity* is the fraction of adjacent map pixels that are either both in the solvent region or both outside the solvent region. This number is not absolute because it depends on the grid intervals of the map, but in general it should be greater than 0.9. This is the only place in which SHELXE divides the map sharply into solvent and not-solvent regions; for the density modification a fuzzy solvent boundary is employed. If one heavy atom enantiomorph gives appreciably larger values for both parameters, it is the correct one and the other SHELXE job can be killed. If the values are very similar for the two jobs it

means either (a) the substructure sites correspond to a centrosymmetric structure (e.g. one site in space group $P2_1$) in which case both SHELXE jobs should produce correct phases (except for SIR in which case a double image will be produced), or more likely (b) the substructure solution is incorrect. A large divergence in the *contrast* and *connectivity* values for the two heavy atom enantiomorphs is a good indication that the map (corresponding to the higher values) will be good. The *contrast* and *connectivity* values also serve to indicate whether the density modification has converged or not: if not, it may be worth increasing the number of cycles (e.g. **`-m20`** instead of **`-m10`**).

SHELXE lists the parameters used at the start of the output, but generally no other changes will be needed. In general after SHELXE has been run in this way, density modification - if possible taking NCS into account - should be performed using programs such as DM, SOLOMON or RESOLVE. Since SHELXE uses different algorithms to these programs and tries to estimate realistic weights (fom), it will usually be a good idea to perform say 10 cycles density modification in SHELXE before performing more sophisticated density modification with these programs. If necessary density modification can be switched off in SHELXE simply by leaving out the **`-m`** flag. In exceptional cases involving very high resolution (better than 1.5Å) or high solvent content (>60%) the program can produce rather high quality maps; this is illustrated by the examples below. In such cases it is worth doing more density modification cycles (20-50).

The additional flag **`-b`** enables the final phases to be applied backwards to generate a special Fourier showing the anomalous sites only. These phases are written to *name.pha* and the corresponding *.res* file (which could in theory be recycled through SHELXE) to *name.hat*. If the structure was inverted then these will correspond to the inverted heavy atoms and space group too.

Phasing may be started from a phase file (XtalView *.phs* format or *.fcf* from SHELXL LIST 6) , in which case this filename should be specified in full and must have the extension *.phi* or *.fcf*; only one file should be specified, and some flags like **`-h`** and **`-i`** cannot be used, e.g.

```
shelxe junk.phi -m8 -s0.4
```

Note that the *junk.phs* file from a previous SHELXE run has been renamed as *junk.phi* here, and a new *junk.phs* file will be created.

The other SHELXE parameters can usually safely be left at their default values; further details may be found below.

Examples of MAD and SAD phasing with SHELXD/SHELXE

Two tests are provided to test SHELXD and SHELXE in combination. The files may be downloaded from the *egs* subdirectory at the SHELX ftp site. The first is based on four-wavelength MAD data and was kindly provided by Zbigniew Dauter. Further details may be found in Li et al. (2000); the PDB code is 1C8U. There are a total of eight unique selenium sites exhibiting twofold NCS. The 2.5Å MAD data have been processed using XPREP to create the file *jia-fa.hkl* that contains *h, k, l,* F_A, $\sigma(F_A)$ and α in format 3I4,2F8.2,I4 where α is the estimated angle in degrees that has to be added to the heavy atom phase to give the protein phase (α is needed for SHELXE but not SHELXD). The native data in the SHELX **`HKLF 4`** format file *jia.hkl* extend (optimistically) to 1.9Å; these are required for SHELXE but not SHELXD. The instruction file *jia-fa.ins* for SHELXD (also created using XPREP, but edited to add the **`NTRY 10`** instruction) is as follows:

```
TITL jia-fa in C222(1)
CELL 0.98000  96.0000 120.0000 166.1300  90.000  90.000  90.000
ZERR   16.00   0.0192   0.0240   0.0332   0.000   0.000   0.000
LATT -7
SYMM -X, -Y, 0.5+Z
SYMM -X, Y, 0.5-Z
SYMM X, -Y, -Z
SFAC SE
UNIT 128
PATS
FIND 8
MIND -3.5
NTRY 10
HKLF 3
END
```

This requests SHELXD to search for 8 sites (**`FIND 8`**) at least 3.5Å apart from each other by dual-space recycling with Patterson seeding. The space group is $C222_1$. The program is started by:

```
shelxd jia-fa
```

(or **`xm jia-fa`** using the Bruker SHELXTL system) which requires that the SHELXD executable is in the PATH. The program selects 3663 E > 1.5 for the dual-space recycling and reports a mean $|E^2-1|$ of 0.778, typical of the non-centrosymmetric distribution expected for MAD F_A values (SAD or SIR should give a centrosymmetric value close to 0.968, the expected non-centrosymmetric value is 0.736). As this substructure is not difficult to solve (it was originally solved with SHELXS!) more than half of the 10 attempts give correct solutions with CC values of about 35.4 and 26.9% and a PATFOM of about 16.6. These clearly identify the correct solutions, as do the other figures of merit, albeit less decisively.

Examination of the *jia-fa.res* or *jia-fa.pdb* output file shows a big drop in the peak-heights between the eighth (correct) peak and the ninth (noise) peak. Even if the number of sites specified with **FIND** is not exactly 8, the occupancy refinement (set by **TANG -0.9** which is the default for substructure solution) ensures that the eight sites have much higher occupancies (≥0.78) than the rest (≤0.15). Since the number of peaks found could still have a slight bias towards the number specified on the **FIND** instruction it may be worth experimenting with different numbers. The heavy atom substructure can be displayed with XP or RASMOL (deleting the noise peaks first and using the ball and stick option); this enables the two-fold NCS axis to be seen. Alternatively the crossword tables given in the *jia-fa.lst* file for the correct solutions can be interpreted to give two groups each containing four peaks with similar distances within each group. I

SHELXE can now be used to calculate phases from the heavy atom coordinates in the *jia-fa.res* file. It is always necessary to test both heavy atom enantiomorphs because the hand of the best SHELXD substructure solution may equally well be correct or inverted (because SHELXD uses random numbers). Note that in general it is not even necessary to look at the *.res* file or delete noise peaks, the occupancy refinement in SHELXD ensures that the noise peaks have very low weights. The two command lines for the SHELXE jobs are:

```
shelxe jia jia-fa -m10 -s0.66
```

```
shelxe jia jia-fa -m10 -s0.66 -i
```

Ten cycles density modification are to be performed (**-m10**) and the solvent content is 66%. The simplest way to estimate the solvent content is to assume that the average amino-acid has a volume of 140Å^3. The second job also inverts the substructure found by SHELXD and stored in the file *jia-fa.res* (if necessary SHELXE would invert the space group too, but that is not needed here). Note that the **-h** switch is not used because the selenium atoms were not present in the native data (file *jia.hkl*) that were collected on the corresponding sulfur-containing native protein to a higher resolution that was possible with the selenium-containing crystals.

The first number to check is the correlation coefficient between E_c (calculated from the substructure) and E_o from the data in the *jia-fa.hkl* file. A very low (<10%) or even negative *CC* would indicate that something is seriously wrong. Note however that the full resolution range is used for this calculation, so the value will often be different from that reported by SHELXD. Here the *CC* is about 30% which is fine. The program uses the resolution dependence of this *CC* to estimate *sigma-A* weights (Read, 1986) for the initial centroid phases that are obtained from the Harker construction (Harker, 1956). At this point both substructure enantiomorphs are still equally likely. During the density modification, a clear

distinction develops between the two enantiomorphs because only one has the sort of electron density distribution expected for a protein. In the first cycle, the values for the *contrast* and *connectivity* are about 0.028 and 0.716 resp. for the false enantiomorph and 0.057 and 0.841 for the correct one; at the end the corresponding values are 0.183 and 0.893 (false) and 0.896 and 0.958 (correct). During the density modification the mean weight is normalized to 0.3 and at the end the program uses an empirical method to estimate absolute weight; a side effect of this is that the final *contrast* and *connectivity* are a little different from the values during the density modification. Since the contrast has not quite converged after 10 cycles it might be better to do a few more. The file *jia.phs* or *jia_i.phs* corresponding to the correct enantiomorph may now be inspected with e.g. XtalView or improved futher with DM, SOLOMON or RESOLVE. This might also enable the two-fold NCS to be exploited (obtaining the necessary transformation from the heavy atom sites). In this case this is scarcely necessary, the mean map correlation coefficient with the structure deposited in the PDB is about 0.83 and increases to about 0.88 when **-m30** is used instead of **-m10**, so autotracing could be attempted directly with wARP (the native data extend to 1.9Å). At the end of the SHELXE run the program tries to estimate the map correlation coefficient as a function of the resolution; although this works well in this example it should be regarded as very preliminary, in other cases the values can be significantly overestimated. The reasons for the high quality of the resulting map are the excellent quality of the data and the rather high solvent content.

For further details of MAD substructure solution using XPREP and SHELXD Thomas Schneider's tutorial should be consulted (http://shelx.uni-ac.gwdg.de/~trs/mad/mad.html).

The second macromolecular example involves the phasing of the orthorhombic modification of bovine trypsin using the weak anomalous signal from the one calcium and fourteen sulfur atoms in the structure. The data were collected to 1.2Å on a Bruker SMART6000 system and processed with SAINT, SADABS and XPREP. The *try2-an.ins* file is as prepared by XPREP except that the resolution has been truncated to 1.7Å using the **SHEL** instruction (as suggested by XPREP), **NTRY 10** has been added to prevent overkill, and the minimum interatomic distance has been set to 1.7Å to accommodate S-S bonds.

```
TITL try2-an in P2(1)2(1)2(1)
CELL 1.54178  53.9001  56.9556  66.0552  90.000  90.000  90.000
ZERR    8.00   0.0013   0.0017   0.0019   0.000   0.000   0.000
LATT -1
SYMM 0.5-X, -Y, 0.5+Z
SYMM -X, 0.5+Y, 0.5-Z
SYMM 0.5+X, 0.5-Y, -Z
SFAC S
UNIT 400
SHEL 999 1.7
PATS
```

```
FIND 15
NTRY 10
MIND -1.7
HKLF 3
END
```

The program reports a mean $|E^2-1|$ of 0.930 - confirming the centrosymmetric statistics expected for SAD data - and selects 2832 $E > 1.5$. Several of the ten attempts lead to solutions with *CC* values of about 37.8 and 22.0 and *PATFOM* of about 8.2. There is a break after the first peak (calcium) with height 1.00 and the second (sulfur, 0.48) and then a fall-off at peaks 15-18 (0.31, 0.29, 0.19, 0.09). In fact the unexpected 16th and 17th peaks are the two sulfate ions that have higher B-values than the other sulfurs. XP or RASMOL reveal the presence of 6 disulfide bonds, the other two sulfurs are methionines. The relatively high success ratio is partly due to the presence of the calcium atom.

The two SHELXE jobs for the two substructure enantiomorphs are now started as follows:

```
shelxe try2 try2-an -h -m10 -s0.38
```

```
shelxe try2 try2-an -h -m10 -s0.38
```

The solvent content is only 38% but the substructure atoms are part of the structure corresponding to the native data (*try2.hkl*) because all the data come from the same data collection on the same crystal. The *CC* for the substructure (15%) indicates that the anomalous signal is not very strong, and the separation of the two enantiomorphs (*contrast* and *connectivity* start at 0.156 and 0.701 resp. for the wrong enantiomorph and at 0.150 and 0.791 for the correct one, at the end the corresponding values are 0.343 and 0.917 (false) and 0.402 and 0.939 (correct). The mean weights after resolving the two-fold ambiguity and after including the contributions from the anomalous scatterers also clearly define the correct enantiomorph. The mean map correlation coefficient with the final refined structure is 0.83, and this can be increased to 0.94 by running SHELXE with **`-m30 -w0`**. The switch **`-w0`** leads to the gradual elimination of the original phase information during density modification, and is only effective for very high resolution data; it should be used with care because in less favorable cases it can lead to a deterioration of the phases. The reasons for the remarkable quality of the final maps from this SAD experiment are the high data quality (the redundancy was 16.3 to 1.2Å) and the very high resolution.

For a recent study of the feasibility of solving trypsin by means of the anomalous scattering of the calcium and sulfur atoms see Yang & Pflugraph (2001).

SHELXD / XM instructions

SHELXD is started with the command line:

shelxd *name*

(similarly **xm *name***) and expects to find both input files *name.ins* and *name.hkl* in the current directory. It writes a summary to the current window (standard output) and creates the files *name.lst* (more extensive listing file) and *name.res* (SHELX format atoms, crystal coordinates).

The following instructions may be included in the *.ins* file. Default values are given in square brackets; the # sign indicates that the default depends on other instructions:

TITL, **CELL**, **ZERR**, **LATT**, **SYMM**, **SFAC** and **UNIT** as usual (see the SHELX manual).

TRIC (or TRIK)

Flags expansion to non-centrosymmetric triclinic for all calculations.

SHEL *dmax* [infinity], *dmin* [0]

Resolution limits in Å for all calculations. Both limits must be specified but it does not matter which is given first.

NTRY *ntry* [0]

Number of global tries if starting from random atoms, **PATS** or **GROP**. If *ntry* is zero or absent, the program runs until it is interrupted by writing a *name.fin* file in the current working directory.

PATS +*np* or -*dis* [100], *npt* [#], *nf* [5]

Calculates and stores Patterson. Using top *np* peaks or a random orientation vector of length

|*dis*|, tries *npt* random translations, selecting the one with the best Patterson minimum function PMF (see **PSMF**). When selecting a vector from the list of unique Patterson peaks, special vectors are ignored and the highest vector is chosen from *nf* random selections. This favors the highest peaks but (if *nf* is not too large) also allows lower peaks a chance. For examples, with the default *np* = 100 and *nf* = 5, the chance is 39.5% that one of the first 10 vectors will be chosen and 91.9% that one of the first 50 will be chosen. The default value of *npt* is 9999 for space groups with a floating origin and 99999 for other space groups. When the space group is *P*1, an extra atom is placed on the origin in addition to the two-atom vector employed for the translation search. In the special case when **FIND 1** is specified with **PATS**, a single atom random translation search is performed instead of using a vector.

If the first parameter is negative, *nf* random oriented vectors of length |*dis*| are compared on the basis of their heights in the Patterson and the 'best' used for the translation search.

If **PATS** is used together with a second **FIND** parameter *ncy* greater than zero (or **FIND** followed by only one number) a full-symmetry Patterson superposition minimum function (i.e. a superposition based on the two peaks and all their symmetry equivalents) is used to locate the starting atoms for the first **FIND** cycle. **PATS** and **GROP** are mutually exclusive.

GROP ***nor*** **[99],** ***E*****$_g$** **[1.5],** ***d*****$_g$** **[1.2],** ***ntr*** **[99]**

The dual-space direct methods is seeded by a 6D search for small rigid group to find a high value (not necessarily the global maximum) of $\Sigma E_c^2(E_o^2-1)$ for the reflections with $E > E_g$ and $d > d_g$, where *d* is the resolution in Å. For each of *nor* random orientations, the local maxima of this function are found starting from *ntr* random translations, and the atom positions corresponding to the orientation/translation combination that gives the highest value for this function are used to initiate the dual-space recycling.

The search model is read from PDB-format **ATOM** or **HETATM** records in the *.ins* file. All other PDB records should be removed. The atomic number is deduced from the atom name applying PDB rules. A short piece of alpha-helix might be used for solving small proteins and a diglucose fragment might be suitable for cyclodextrins. In practice, a thorough six-dimensional search (with a large *nor* value and $E_g = 0$) using **GROP** is rather slow, but when used in combination with **TRIK, GROP** is much faster because then only a three-dimensional search is required.

PSMF ***pres*** **[3.0],** ***psfac*** **[0.34]**

pres is the resolution of the Patterson in terms of minimum ratio of the number of grid points along an axis and the maximum reflection index along that axis. If *nres* is negative a 'super-sharp' Patterson with coefficients $\sqrt{(E^3F)}$ is calculated (in which case a finer grid is advisable, i.e. **PSMF -4**), otherwise a normal F^2 Patterson is used. *psfac* is the fraction of the lowest values in the sorted list of Patterson heights that is summed to get the PMF.

FRES *res* [3.0]

Resolution of all Fourier syntheses (including the PSMF but excluding the Patterson itself) in terms of the minimum ratio of the number of grid points along an axis and the maximum reflection index used along that axis.

ESEL *Emin* [#], *dlim* [1.0]

Minimum E and high-resolution limit for FIND. The E^2 values are normalized to 1 in resolution shells, then smoothed. *Emin* defaults to 1.2 for *ab initio* structure solution and to 1.5 for heavy atom location (the appropriate value is set as default depending on whether a **PLOP** instruction is present or not).

FIND *na* [0], *ncy* [#]

Search for *na* atoms in *ncy* dual space cycles. If **WEED** is employed, *na* is the number of atoms remaining after the random omit. *ncy* defaults to the largest of (20 or *na*) or, if **PATS** is used, to the smaller of (3*na* and 20). If **FIND** is absent, **PLOP** expands directly from the starting atoms.

TANG *ftan* [0.9], *fex* [0.4]

Fraction |*ftan*| of the *ncy* dual space (**FIND**) cycles are performed using the tangent formula, the rest using a Sim-weighted E-map. *fex* is the fraction of reflections with the largest E_{calc} values to hold fixed when doing tangent expansion to find the remaining phases. **WEED** is only applied to the first |*ftan*|·*ncy* cycles. If *ftan* is negative, the occupancies are refined for the final (1-|*ftan*|)·*ncy* cycles. This is particularly useful for the anomalous sites in *halide soak* experiments, since these often have partial occupancies, but for other substructure

problems it also provides a good check as to how many heavy atom sites are present. It is not recommended for normal *ab initio* applications of SHELXD because the algorithm employed uses a large amount of memory (in the interests of speed).

NTPR ***ntpr*** **[100]**

Maximum number of (largest) TPR (triple phase relations) per reflection. If *ntpr* is negative, E is replaced by $E/[1+\sigma^2(E)]$ in the estimation of probabilities involved in the tangent formula and minimal function, as recommended by Giacovazzo (2001).

MIND ***mdis*** **[1.0],** ***mdeq*** **[2.2]**

|*mdis*| is the shortest distance allowed between atoms for **PATS** and **FIND**. If *mdis* is negative *PATFOM* is calculated, and the *crossword table* for the best PATFOM value so far is output to the *.lst* file. In this case the solution is passed on to the **PLOP** stage if either the CC is the best so far or the *PATFOM* is the best so far. *mdeq* is the minimum distance between symmetry equivalents for **FIND** (for **PATS** the |*mdis*| distance is used). Thus the default setting of *mdeq* prevents **FIND** from placing atoms on special positions. This is usually desirable because it helps to avoid pseudo-solutions such as the 'uraninum atom solution' that are incorrect but fit the tangent formula, but it might be better to change this setting to -0.1 to allow special positions; especially for the location of heavy atom sites obtained by (halide) soaking. For **PLOP** the **PREJ** instruction can be used to control whether peaks on special positions are selected.

SKIP ***min2*** **[0.5]**

During **FIND**, if the second peak height is less than *min2* times the first, the first peak is rejected (before applying **WEED** to reject other peaks). This is sometimes useful to suppress 'uranium atom' solutions. For large equal-atom structures in space group *P*1 where there is a danger of an uranium-atom pseudo-solution it might be a good idea to specify **SKIP 0.99** so that the first peak is ALWAYS rejected!

WEED ***fr*** **[0.3]**

Randomly omit fraction *fr* of the atoms in the dual space recycling (except in the last cycle and the cycles for which no tangent refinement is performed - see **TANG**). **WEED** not applied to the **PLOP** stage.

CCWT *g* [0.1]

All correlation coefficients (*CC*) are calculated using weights $w = 1/[1+g\sigma^2(E)]$. If the $\sigma(E)$ values read from the *.hkl* file are known to be very unreliable, it might be better to set *g* to zero. If XPREP was used to create the file the default value of 0.1 should never need to be changed. The correlation coefficients between E_c and E_o are calculated using the formula:

$$CC = 100\,[\Sigma wE_oE_c\Sigma w - \Sigma wE_o\Sigma wE_c] \,/\, \{\,[\Sigma wE_o^2\Sigma w - (\Sigma wE_o)^2]\,[\Sigma wE_c^2\Sigma w - (\Sigma wE_c)^2]\,\}^{½}$$

TEST *CCmin* [#], *delCC* [#]

If **PLOP** has not yet been entered, the program goes on to the calculation of *PATFOM* and the crossword table (if the first **MIND** parameter is negative) or directly to the **PLOP** stage if the *CC* after dual space recycling is greater than *CCmin*, otherwise the next dual space attempt begins immediately with new starting atoms. *CCmin* is reduced by 0.1% each cycle until a solution passes this test. After **PLOP** has been entered at least once, subsequent attempts go on to *PATFOM* and/or **PLOP** if *CC* is within *delCC* of best *CC* value so far.

If *PATFOM* is calculated, then only solutions with either the best initial (i.e. after the dual space recycling) *CC* so far or the best *PATFOM* so far go on to the **PLOP** stage. Whether or not *PATFOM* is calculated, if **PLOP** is absent the heavy atom sites with the best initial *CC* so far are written to the *.res* and *.pdb* files. If **PLOP** is specified, then the *.res* and *.pdb* files are written after the **PLOP** stage. Since these files are closed and reopened each time, they may be inspected, e.g. using RASMOL (use ball and stick mode) or the Bruker SHELXTL program XP, without stopping the SHELXD job.

The defaults for *CCmin* and *delCC* are 45 and 1 resp. for *ab initio* solutions, and 10 and 5 resp. for heavy atom location (i.e. when **PLOP** is absent).

KEEP *nh* [0]

Number of (heavy) atoms to retain as fixed atoms during **PLOP** expansion. This will

normally only be used when expanding from starting atoms (**PLOP** without **FIND**, **GROP** or **PATS**).

PLOP followed by up to 10 numbers

PLOP specifies the number of peaks to start with in each cycle of the *peaklist optimization* algorithm of Sheldrick & Gould (1995). Peaks are then eliminated one at a time until either the correlation coefficient cannot be increased any more or 50% of the peaks have been eliminated.

PREJ *maxb* [3], *dsp* [-0.01], *mf* [1]

maxb is the maximum number of bonds to atoms or higher peaks, the peak is deleted if there are more. Peaks are also deleted if they are less than *dsp* from their equivalents (**PLOP** only, **FIND** uses second **MIND** parameter), do not output atoms to final *.res* file if less than *mf* atoms in 'molecule'.

SEED *nrand* [0]

Set random number seed so that exactly the same results are generated if the job is repeated (on an identical computer); each integer *nrand* defines a different sequence of random numbers. If *nrand* is omitted or zero, the seed is randomized so a new sequence is always generated.

MOVE *dx* [0], *dy* [0], *dz* [0], *sign* [1]

Shift following atom coordinates (not ATOM/HETATM). This has the same effect as the **MOVE** instruction for SHELXL.

ATOM and HETATM

PDB format atoms for use by **GROP**.

HKLF ***m***

m = 4 for F^2 in *.hkl* file, m = 3 for F (or F_A or ΔF).

END

SHELXE / XE command line parameters

SHELXE (XE in the Bruker Nonius version) is run by means of a single command line. Usually a substructure solved by SHELXD (XM) and stored in the file *name-df.res* provides the initial phases, in which case the program name should be followed by the first part of the filename of the native data (omitting *.hkl*) followed by the first part of the filename used for the substructure data for SHELXD. Examples are given above. Alternatively the phases may be taken from a file *name.phi* or *name.fcf*, in which case this filename should be given in full and only one filename may be specified. The specified filename must have the extension *.fcf* or *.phi*. If the phases are read from a *.fcf* file (written by SHELXL) the cell and symmetry information are also read from this file; if the phases are read from *name.phi* a SHELX format file *name.ins* must be provided to provide the program with this information. Further control is provided by switches, which may be followed by an integer or decimal number with no intevening white space. Normally only the switches **-h**, **-m**, **-s** and **-i** are employed in routine use of SHELXE. If the switch is followed by ***I*** in the list below, an integer number is required, if it is followed by ***R*** a real is required (the decimal point is however not needed if it is a whole number). Here is the complete list of possible switches:

-f: read *F* rather than F^2 from the native *.hkl* file (SHELX **HKLF 3** format).
-i: invert substructure and if necessary space group.
-n: do **not** resolve twofold ambiguity if SIR or SAD (for test purposes only!).
-h: heavy atoms also present in native.
-b*R*: use modified phases to regenerate substructure (possibly followed by *B* value).
-y: use 272 point sphere instead of 92.
-z: do **not** use sharpening (otherwise used).
-d*R*: high resolution cutoff (not normally needed because weights handle it better).
-m*I*: number of cycles of density modification.
-e*R*: fraction of pixels retained for twofold ambiguity resolution.
-s*R*: solvent fraction.
-c*R*: fraction of pixels in crossover (fuzzy) region.
-g*R*: solvent gamma flipping factor.
-w*R*: weight for retaining initial phase information.
-r*R*: grid size (like **FRES** for SHELXD).
-l*I*: reserve space for reflections (the integer parameter is multiplied by 1000000).

-b and **-l** may also be specified without numerical arguments (in the latter case this causes space to be reserved for only 500000 reflections). SHELXE prints a list of all the settings on starting the program.

If the **-b** switch is used, the final native phases are converted to the substructure phases by

subtracting the phase shifts in the *name-df.hkl* file, and written to the file *name.pha*. This may be viewed using XtalView etc. It should show elongated peaks for disulfide bridges at low resolution (*super-sulfurs*) etc. A list of atoms from a peaksearch of the corresponding map in SHELX *.res* format is written to the file *name.hat*. In principle this may contain extra weak sites not found by SHELXD, and may be renamed (to a *.res* file) and used as input for SHELXE.

SHELXE writes the final native phases to the file *name.phs* suitable for viewing with XtalView or further density modification etc. A listing (similar to that appearing on the console) is written to *name.lst*. If the **`-i`** switch is used the files *name_i.phs* and *name_i.lst* are written instead.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30-42.

Hu, S-H., Gehrmann, J., Guddat, L. W., Alewood, P. F., Craik, D. J. & Martin, J. L. (1996). *Structure* **4,** 417-423.

Li, J., Derewenda, U., Dauter, Z., Smith, S. & Derewenda, Z. (2000). *Nature Struct. Biol.* **7**, 555-559.

Giacovazzo, C. (2001).

Giacovazzo, C. & Siliqi, D. (1997). *Acta Cryst.* A**53**, 789-798.

Harker, D. (1956). *Acta Cryst*. **9**, 1-9.

Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *Science* **259**, 1430-1433.

Miller, R., Gallo, S. M., Khalak, H. G. & Weeks, C. M. (1994). *J. Appl. Cryst*. **27**, 613-621.

Read, R. J. (1986). *Acta Cryst*. A**42**, 140-149.

Sheldrick, G. M. & Gould, R. O. (1995). *Acta Cryst.* B**51**, 423-431.

Sheldrick, G. M., Hauptman, H. A., Weeks, C. M., Miller, R. & Usón, I. (2001). *International Tables for Crystallography*, vol. **F**. Edited by E. Arnold, & M. Rossmann, pp. 333-351. Dordrecht: Kluwer Academic Publishers.

Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 501-505.

Woolfson, M. M.

Yang, C. & Pflugraph, J. W. (2001). *Acta Cryst.* D**57**, 1480-1490.