Chapter 5.4

SHELXD

Ab Initio Structure Solution By Patterson and Direct Methods

George W. Sheldrick
Dept. of Structural Chemistry
Tammannstraße 4
37077 Göttingen, Germany
email: gsheldr@shelx.uni-ac.gwdg.de

5.4.1 Introduction to SHELXD

The structure solution program SHELXD 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.

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*. 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:
1. Input atoms in normal SHELX format for expansion using **PLOP**
2. **PATS** for Patterson seeding of the dual-space direct methods
3. **GROP** and a PDB-format model for fragment seeding
4. 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 programs such as 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.

5.4.2 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 http 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). **TWIN** and **BASF** are only applied at the **PLOP** stage, and are ignored by **PATS**, **GROP** and **FIND**.

SHELXD instructions

SHELXD is primarily used for macromolecular structure solution. This manual describes how it may be used for smaller systems. SHELXD is started with the command line: **shelxd *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 sixdimensional 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 'supersharp' 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

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.