Intrinsic phasing

(also known as XT or SHELXT)

MIT, February 22nd 2014

George M. Sheldrick, *Göttingen University*

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

Introduction

SHELXT is a program for solving relatively routine small molecule structures, determining the space group and structure together. Most of the methods it uses are well established, but they are combined in a way that is designed to be robust, efficient and very easy to use.

SHELXT reads standard SHELX format *name.ins* and *name.hkl* files. The CELL, LATT, SYMM, SFAC and HKLF cards are read but the input space group and the rest of the *.ins* file are ignored. The Laue group is extracted but may be overridden with the –L command line switch, e.g. –L15 to make the program try all trigonal and hexagonal Laue groups. Other command line options are possible but are rarely needed, for a full list run SHELXT without a filename. SHELXT is a stand-alone executable, no other files, DLLs, environment variables etc. are needed. To solve a structure enter:

shelxt name

Space groups for small and macro-molecules

For data collection and scaling, we must first establish the Laue group and lattice type. To keep it simple, let us assume that we have an orthorhombic crystal, i.e. Laue group mmm, and a primitive lattice.

If it is a protein, there will be 4 possible space groups – $P2_12_12_1$ (23.9%), $P2_12_12$ (6.1%), $P222_1$ (0.04%) and P222 (0.01%). Note the very unequal distribution in the PDB! We can simply ask a MR program such as PHASER to try all 8 (!) possibilities, and if only one leads to a solution, then it is the correct space group and orientation. Since three-quarters of proteins are solved by MR anyway, this is a popular and reliable method of determining their space groups.

If it is a small molecule, there are 30 orthorhombic-P space groups and 111 different possible combinations of space group and orientation for them, so it is more interesting!

Small molecule example: $[Co(OH_2)_6]^{2+}\,[NH_4]^{+}\,[PO_4]^{3-}$

This is an example taken from Peter Müller's book. XPREP looks for the highest metric symmetry, in this case orthorhombic P, and sets a<b<c. The low R(sym) confirms Laue symmetry mmm.

```
Determination of reduced (Niggli) cell

Transformation from original cell (HKLF-matrix):
  -1.0000  0.0000  0.0000    0.0000  0.0000 -1.0000    0.0000 -1.0000  0.0000

Unitcell:       6.112    6.941   11.196   90.00   90.00   90.00

Niggli form:    a.a =     37.36      b.b =     48.18      c.c =    125.35
                b.c =      0.00      a.c =      0.00      a.b =      0.00

Search for higher METRIC symmetry
Identical indices and Friedel opposites combined before calculating R(sym)

-------------------------------------------------------------------------------
Option A: FOM = 0.000 deg.   ORTHORHOMBIC P-lattice   R(sym) = 0.022 [   1607]
Cell:    6.112   6.941  11.196   90.00   90.00   90.00    Volume:       474.97
Matrix:-1.0000  0.0000  0.0000  0.0000  0.0000 -1.0000  0.0000 -1.0000  0.0000
-------------------------------------------------------------------------------
Option B retains original cell

Select option [A]:
```

XPREP space group determination

XPREP looks for a space group that fits the Laue group, lattice type, systematic absences, intensity statistics and frequency of the space group in the CSD. These last two factors have lower weight in CFOM. A and B have different axial orientations but the same CFOM, so XPREP cannot distinguish them. $<|E^2-1|>$ indicates non-centrosymmetric (or twinning).

```
Mean |E*E-1| = 0.677 [expected .968 centrosym and .736 non-centrosym]


Systematic absence exceptions:

        b--   c--   n--  21--   -c-   -a-   -n-  -21-   --a   --b   --n  --21

N       508   508   580    22   438   445   429    29   319   316   311    39
N I>3s  205   205     0    19   391   435   384     0   302   276   264     0
<I>   463.4 464.0   3.1 895.3 565.6 621.6 568.7   2.8 892.6 542.6 833.4   2.3
<I/s>  10.4  10.4   0.4  15.3  16.3  19.4  16.4   0.4  17.7  15.8  14.8   0.5


Identical indices and Friedel opposites combined before calculating R(sym)

Option  Space Group  No.  Type  Axes  CSD  R(sym) N(eq)  Syst. Abs.   CFOM

[A] Pmn2(1)        # 31  non-cen  2    53  0.022   1607  0.5 / 10.4   3.20
[B] Pmn2(1)        # 31  non-cen  5    53  0.022   1607  0.5 / 10.4   3.20
[C] Pmmn           # 59  centro   3    42  0.022   1607  0.5 / 10.4  11.80

Select option [B]:
```

Structure solution with XT/SHELXT

SHELXT finds *one* solution. Both R1 and α (normalized mean square phase error) are convincing. The Flack x may improve after anisotropic refinement. All atoms are correctly assigned. Note that a change in cell orientation is also involved. Total time on an i7 desktop was about 1 sec.

```
Try N(iter)  CC    R(weak)   CFOM    best  Sig(min) N(P1) Vol/N
  1   100   93.98  0.0278  0.9120  0.9120  2.422    27   17.59
  2   100   94.27  0.0256  0.9172  0.9172  2.588    27   17.59
  3   100   94.00  0.0278  0.9122  0.9172  1.898    27   17.59
  4   100   94.26  0.0252  0.9174  0.9174  2.507    27   17.59
  5   100   94.20  0.0250  0.9170  0.9174  2.507    27   17.59

5 attempts, solution 4 selected with best CFOM = 0.9174, Alpha0 = 0.333

0 Centrosymmetric and  56 non-centrosymmetric space groups evaluated

  R1  Rweak Alpha   Orientation  Space group Flack_x  File  Formula
0.033 0.016 0.001  a'=c, b'=a, c'=b  Pmn2(1)  0.23  t071_a  O10 P Co N
```

Output file from SHELXT (t071_a.res)

```
REM Solution 1  R1  0.033  Rweak  0.016,  Alpha = 0.0015  in Pmn2(1)
REM Flack x =  0.226 ( 0.012 ) from Parsons' quotients
REM Formula:  O10 P Co N
TITL pmn21b in Pmn2(1)
CELL  0.71073   6.9410   6.1120  11.1960   90.000   90.000   90.000
ZERR    2.00   0.002    0.003    0.002    0.000   0.000   0.000
LATT -1
SYMM 1/2-X, -Y, 1/2+Z
SYMM 1/2+X, -Y, 1/2+Z
SYMM -X, Y, Z
SFAC O H P CO N
UNIT 20 32 2 2 2
L.S. 10
BOND
LIST 6
FMAP 2
PLAN 20                                          Electron count
ANIS                                                    ↓
RIGU
CO01  4  0.50000  0.86717  0.49884 10.50000  0.01068  27.98
P002  3  1.00000  0.49687  0.62660 10.50000  0.01000  16.12
O003  1  1.00000  0.51268  0.49086 10.50000  0.01200   7.76
O004  1  0.70713  0.97162  0.61863 11.00000  0.01877   7.71
O005  1  1.00000  0.72904  0.68051 10.50000  0.01325   7.62
O006  1  0.27539  0.76252  0.38884 11.00000  0.01556   7.60
O007  1  0.81857  0.37423  0.66825 11.00000  0.01396   7.50
O008  1  0.50000  0.54957  0.58052 10.50000  0.01833   7.14
O009  1  0.50000  1.18203  0.41591 10.50000  0.02240   6.92
N00A  5  1.00000  0.12487  0.36067 10.50000  0.01931   5.92
HKLF 4
END
```

The P1 approach

It is well established that most direct methods often work best if the data are first expanded to P1 [Sheldrick & Gould, *Acta Cryst.* B51 (1995) 423-431] . This suggests the following approach:

1. Assuming the Laue group to be known, equivalent intensities are averaged and the data then expanded to P1.
2. The phase problem is solved in P1. The result is an electron density map and the corresponding phases.
3. The *phases* are used to determine the correct space group and the translation necessary for the electron density map to fit it.
4. The phases are averaged in this space group and used to calculate an improved density.
5. The maxima of the density are assigned to atoms.

Starting from the Patterson function

Unlike almost all direct methods of the last three decades, SHELXT is (almost) *deterministic*. Although dual-space direct methods usually start from random phases or atoms, in the presence of some heavier atoms a considerable speed-up can be achieved by starting from a Patterson superposition minimum function.

If two copies of the Patterson, displaced from one another by a vector corresponding to a strong Patterson peak, are superimposed and their minimum function calculated, it should correspond to a double image of the structure with shifted origins in the effective space group P1.

So if we are planning to solve the structure first in P1 anyway, this is an excellent starting point, and should never be worse than starting from random phases.

Random OMIT maps

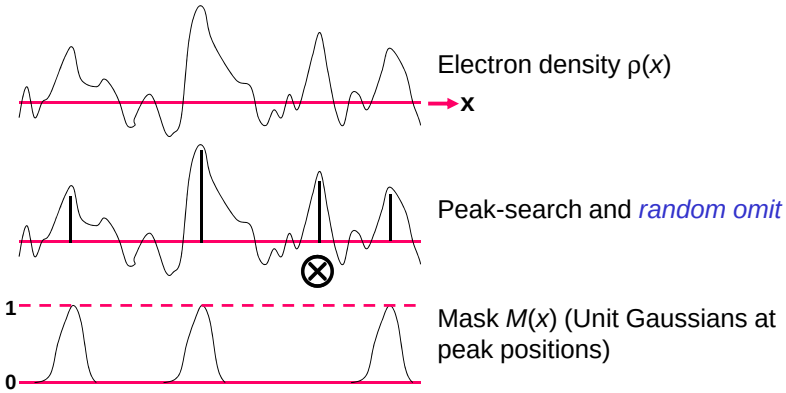
Omit maps were frequently used by protein crystallographers to reduce *model bias* when interpreting unclear regions of a structure. A small part (<10%) of the model is deleted, then the rest of the structure refined (often with simulated annealing to reduce memory effects) and finally a new difference electron density map is calculated.

A key feature of SHELXD when it is used to solve large equal-atom structures is the use of *random omit maps*. About 30% of the peaks are omitted at random each cycle and the remaining atoms used to calculate phases. This can be regarded as a *perturbation* of the density, like charge flipping.

Random omit maps are also employed in the dual-space recycling part of SHELXT (making it somewhat less deterministic).

Random omit without atoms

The following dual space algorithm enables the *random omit* method to be combined efficiently with FFTs in both directions. It achieves similar results to charge flipping, but imposes more atomicity.



The new density is calculated as $\rho'(x) = \rho(x) \cdot M(x)$. This also truncates negative density and sharpens the remaining atoms!

A phase determination strategy

1. Generate starting maps in P1 by twofold and threefold Patterson superpositions. This of necessity produces a trial structure in P1.
2. Iterate using the *random omit mask* perturbation. Use sharpened amplitudes $G_{hkl} = (EF)^{½} \{ F_o^2 / [F_o^2 + \sigma^2(F_o)] \}$ and Fourier coefficients $3G_o - 2G_c$ with phases ϕ_c.
3. The phases with the best CFOM are used to find the space group followed by further density modification. CFOM = 0.01CC – R(weak), where R(weak) is the mean value of E_c^2 for reflections with smallest E_o^2.

Unlike most other direct methods, this algorithm works reasonably well with heavy atoms on special positions and for pseudo-symmetric structures. Because it indirectly assumes *atoms* it is less demanding on data quality and completeness than charge flipping, but charge flipping should work better for modulated and severely disordered structures, because it *doesn't assume atoms*.

Using phases to find the correct origin and space group

PLATON can find the space group and place its origin correctly starting from atom positions in P1. The disadvantage of this approach is that tolerances are required to decide whether atom positions are the same within experimental error in the higher space group.

Giacovazzo [*J. Appl. Cryst* 33 (2000) 307] and Palatinus [*J. Appl. Cryst.* 41 (2008) 975] suggested ways of finding the space group and the required origin shift using only the phases. The two approaches are similar but Giacovazzo finds the full space group directly (but at his own admission, slowly) whereas Palatinus searches for individual symmetry elements and then uses them to construct the space group.

An advantage of this approach is that it is possible to define a single figure of merit to decide which possible space group is more likely.

The phases of equivalent reflections

The phase ϕ_m of the equivalent reflection $\mathbf{h}_m$ is derived from the phase ϕ of the (prime) reflection **h** by:

$$\phi_m = \phi - 2\pi \mathbf{h} t_m = \phi - 2\pi (h t_1 + k t_2 + \ell t_3)$$

For example in $P3_1$: (for symmetry operator #2)

$$h_2 = 0h + 1k + 0\ell = k$$
$$k_2 = -1h - 1k + 0\ell = -h-k$$
$$\ell_2 = 0h + 0k + 1\ell = \ell$$

So $\mathbf{h}_2$ is $k, -h-k, \ell$ with phase:

$$\phi_2 = \phi - 2\pi \mathbf{h} t_2 = \phi - 2\pi (0h + 0k + \tfrac{1}{3}\ell) = \phi - (\tfrac{2}{3})\pi\ell$$

This relation is only valid if the space group is correct and the structure has been shifted so that the origin is correctly placed.

Phases and translation

If the whole structure is shifted by **Δx** we can write: $\mathbf{x}_j' = \mathbf{x}_j + \Delta\mathbf{x}$

Thus $\mathbf{F_h'} = \sum_j f_j \exp(2\pi i\, \mathbf{h} x_j') = \sum_j f_j \exp(2\pi i\, \mathbf{h}(x_j+\Delta x))$

$= [\sum_j f_j \exp(2\pi i\, \mathbf{h} x_j)] \exp(2\pi i\, \mathbf{h}\Delta x)$

$= \mathbf{F_h} \exp(2\pi i\, \mathbf{h}\Delta x)$

i.e. $\phi' = \phi + 2\pi \mathbf{h}\Delta\mathbf{x}$

This provides us with a way to find the origin shift **Δx** using the phases, by fitting the phase differences between equivalent reflections **h'** and **h** to **$2\pi\mathbf{h}\Delta\mathbf{x}$.** This calculation has to be performed modulo 2π!

The α figure of merit

In my hands both the Giacovazzo and Palatinus figures of merit performed well. I am using the α value defined by Palatinus because it has a more direct physical meaning (the normalized mean square phase error). For a prime reflection **h** and a symmetry equivalent $\mathbf{h}_m$ we define:

$$q = \{ \phi_m - \phi + 2\pi[\mathbf{h} t_m + \Delta\mathbf{x}(\mathbf{h}_m - \mathbf{h})] \} \text{ modulo } 2\pi$$

For the correct origin shift Δx and the correct space group q should be close to zero. An F^2-weighted sum of q^2 over all pairs of equivalents for all reflections, normalized so that it would have a value of 1.0 for random phases, is then the figure of merit α.

Finding the origin in a centrosymmetric space group

Only when the inversion center of a centrosymmetric structure is at the origin will the phases be 0 or π, otherwise they have general values. To find the origin shift needed to bring an inversion center to the origin, we can double the phases (so that they should all be zero if the origin is correct) and then perform a Fourier transformation:

$$P_X = \sum_h |F_h|^2 \exp(-2\pi i \cdot 2\phi_h) \exp(-2\pi i \cdot hX)$$

The electron density should then be shifted by **X/2** to bring it to the true origin, where **X** is the position of the maximum of this *origin shifted Patterson function*. In SHELXT the α figure of merit for this P1 to $P\bar{1}$ conversion is referred to as α_0. It should be less than about 0.25 for a centrosymmetric space group.

It is still necessary to take into account that not all inversion centers are equivalent in all space groups. However it is still much faster to test all possible non-equivalent inversion centers than to do a full 3D grid search to find the best value of α.

Calculating α in non-centrosymmetric space groups

A 3D grid search for α would be slow because it is not suitable for a FFT, so it is divided into 2D and 1D searches. All non-centrosymmetric space groups in the given Laue group are tested as follows, taking axis transformations to obtain conventional settings into account where necessary.

1. For P1 no search is required, α is undefined (set to zero).
2. For the space groups Pm, Pc, Pn, Cm and Cc a 1D line search is performed.
3. For all other polar space groups, a 2D grid search is performed.
4. For all other space groups, a 2D grid search is followed by a 1D line search.

In all cases (including centrosymmetric) the Δx value obtained by interpolation is then refined further to minimize α.

Preliminary element assignment

In each of the possible space groups, if necessary after re-orientating the axes to obtain the conventional setting, further dual-space recycling is performed to improve the quality of the electron density. The peaks in this map are integrated to get electron counts and these are used to assign atoms, assuming that all possible elements present have been specified (SFAC is used but not UNIT).

There is a problem in putting the electron counts onto an absolute scale. Currently this is solved by looking for some typical organic junk and assuming that it is carbon, or for inorganics by looking for typical groups such as oxyanions. If this cannot be done, the program assumes that the atom with the highest density corresponds to the element with the highest atomic number on the SFAC instruction.

At this point some simple chemical rules are applied to avoid nonsense assignments. If a heavy atom is clearly present but not given on SFAC, the program suggests Br or I.

The free lunch algorithm

If the data are incomplete, SHELXT can simply invent the data that it would have liked to have but doesn't (the *free lunch algorithm*, Caliandro *et al., Acta Cryst.* (2005) **D**61 556-565). This can be useful in the following cases :

1. It was not possible to measure complete data, e.g. because a high-pressure cell was used.

2. To check what the compound is (and possible reject the crystal) before data collection is complete.

3. To obtain more complete structures of poorly diffracting crystals by artificially extending the resolution.

4. The electron density integration used in the element assignment works better if strong but missing low order reflections are included in this way.

Isotropic refinement and the Flack x parameter

After assigning the atoms, an isotropic refinement is performed and R1 calculated. Atoms with very high U-values are eliminated after the isotropic refinement but to save time no further refinement is performed. This may well change in future versions.

For non-centrosymmetric space groups, a Flack x parameter is estimated by the Parsons' quotient algorithm. If x is greater than 0.5 the coordinates and if necessary the space group are inverted. *Thus the structure determined by SHELXT almost always has the correct hand!*

x can also be good indication as to whether the space group is correct. However the value so obtained is not as good as the value after the final refinement.

Bringing the atoms together

The UNIQ instruction in XP was the standard for building molecules, but the following algorithm used in SHELXT is better, because it does not require that elements and hence covalent radii are correctly assigned to the atoms!

1. Generate the SDM (Shortest Distance Matrix – shortest distances between unique atoms, taking symmetry into account).
2. Set a flag to −1 for each unique atom, then change it to +1 for one atom - it does not matter which.
3. Search the SDM for the shortest distance for which the product of the two flags is −1; *if none, exit*.
4. Symmetry transform the atom with flag −1 for this distance so that it is as close as possible to the atom with flag +1, then change its flag to +1.
5. GOTO 3

This diabolically simple algorithm not only builds the molecules as we would intuitively expect them whatever the space group, but also clusters them in a chemically sensible way, making the structure instantly recognizable.

Pseudosymmetry

When the space group is centrosymmetric but the heavy atom substructure is non-centrosymmetric, there can be many apparent solutions. Sometimes the result is a mess. For this Pt complex, R1 and x indicate the correct model.

R1	Rweak	Alpha	Space group	Flack_x	File	Formula
0.634	0.338	0.051	P4/mmm		t240_a	C35 N6 Pt96 I
0.203	0.004	0.069	P4(2)/mmc		t240_b	C42 N12 F6 Cl8 Pt
0.637	0.398	0.070	P4(2)/mnm		t240_c	C22 N8 Pt46 I
0.208	0.007	0.071	P4/mnc		t240_d	C38 N26 F14 Cl4 Pt
0.552	0.137	0.035	P-4m2	0.48	t240_e	C59 N2 F8 Cl10 Pt67
0.165	0.004	0.051	P4mm	0.50	t240_f	C83 N8 F25 Cl16 Pt2
0.168	0.004	0.053	P422	0.50	t240_g	C39 N16 F14 Cl5 Pt
0.181	0.004	0.053	P-42m	0.49	t240_h	C46 N16 F44 Cl12 Pt I
0.146	0.003	0.054	P4(2)mc	0.49	t240_i	C41 N5 F5 Cl2 Pt
0.123	0.003	0.055	P4(2)2(1)2	0.09	t240_j	C30 N6 F14 Cl2 Pt
0.139	0.003	0.056	P-42(1)c	0.47	t240_k	C26 N16 F12 Pt
0.157	0.004	0.072	P-42(1)m	0.49	t240_l	C41 N14 F12 Cl3 Pt
0.515	0.120	0.072	P-42c	0.46	t240_m	C16 Pt68 I
0.529	0.146	0.073	P4(2)22	0.47	t240_n	C20 N4 Pt58 I
0.166	0.004	0.073	P42(1)2	0.44	t240_o	C48 N21 F5 Cl4 Pt
0.142	0.006	0.074	P4(2)nm	0.50	t240_p	C42 N22 F12 Cl2 Pt
0.192	0.007	0.075	P4nc	0.50	t240_q	C41 N22 F14 Cl9 I
0.164	0.007	0.076	P-4n2	0.49	t240_r	C42 N22 F14 Cl4 Pt

Command line options

If SHELXT started without a filename it prints a list of possible command line options. –L is useful to override the Laue group specified by the SYMM cards, e.g. –L15 to try all trigonal and hexagonal space groups.

–a causes the program to try all space groups in the given Laue group instead of stopping after the first plausible solution and is the first thing to try if the P1 solution appears successful but no suitable space group has been found. In extreme cases –a0.5 may be tried.

If the figures of merit in the P1 stage are poor, -i2 or –i4 may be better than the default –i3. This varies the coefficients for the Fourier maps. Also –u1.4 (tangent formula for E>1.4) or –m500 (5 times as many tries) may be worth trying. For very tightly packed structures (e.g. some borides) the volume per atom should be reduced (e.g. –v8 rather than the default –v13).

If the resolution of the data is poor, it may be worth varying –d and –e from the values used by the program.

Weak points in the current version

The current version obtains fully correct structures for about 50% of the structures tested and most of the rest have the right atoms but some of them are wrongly assigned (typically N/C and S/P). The space group is found correctly in at least 97% of cases, although some of them had defeated XPREP. SHELXT showed that several of the ca. 630 test structures had been refined in the wrong space group!

Most of the small number of failures involved severe pseudo-symmetry or disorder, especially 'whole molecule disorder', or twinning, violating the assumption that the structure consists of ***atoms***. Non-centrosymmetric structures with a centrosymmetric arrangement of very heavy atoms can also be problematic.

The largest structure solved and assigned completely correctly so far had 360 unique atoms in $P2_1/n$, but it was a particularly favorable case. Usually SHELXD is more effective for very large or twinned structures.

Changes in version 2014/1 and 2014/2

SHELXT is now fully parallel and the code has been hand-optimized. It now takes a couple of hours rather than days for my set of 630 test structures.

-L16 tries monoclinic with the current a axis as the b axis, -L17 does the same for c. This can be useful when β is accidentally very close to 90°.

The option –u1.4 uses the tangent formula for E-values greater than 1.4 in the P1 solution part. The default –u99 switches It off. The tangent formula can be useful for large equal-atom structures, but in general it is better not to use it.

-o switches off the Patterson seeding (not recommended).

The P1 solutions are sorted on CFOM = 0.01CC – g.R(weak). g is set using –j. To use just CC, input –j0 on the command line.

If an atom heavier than scandium is expected (SFAC), -a is set automatically.

Future plans

SHELXT is currently at the beta-test stage and improvements are still being made; the feedback from beta-testers is essential for this. Most of the beta-testers are Bruker users and the Bruker email forum proves very useful for discussing new features. The beta-test has an expiry date because I do not want to have to support older buggy versions for the rest of my life.

The plan is to release SHELXT as part of the standard SHELX system in good time for the IUCr Meeting in Montreal in August 2014. It will then no longer have an expiry date and will be accessible and documented via the SHELX homepage. SHELXT will also play a part in the SHELX Workshop on the first day of the Meeting (August 5th).

Future plans include a better treatment of pseudo-symmetry and improved element assignment, as well as the automatic solution of twinned structures. This may take some time.

It is to be expected that SHELXT will make SHELXS obsolete but not SHELXD (still better for twins and very large equal-atom structures).

Acknowledgements

I am very grateful to all the beta-testers for their helpful suggestions and test data.