SHELX results for phasing and refinement workshop ACA2011 New Orleans – George Sheldrick

2qvo weak anomalous test

There is an excellent presentation of the XDS data processing and SHELXC/D/E phasing for this test by Kay Diederichs http://strucbio.biologie.uni-konstanz.de/xdswiki/index.php/2QVO so there is little to add. Kay processed the two runs separately, we will call them 2qvo-1 and 2qvo-2. Merging the two together (with XPREP) gives 2qvo-12. In addition James Holton processed the two runs together with ELVES and there is another link to a dataset provided by the organisers that turns out to be the same ELVES data. We will call the merged ELVES data 2qvo-el.

In the space group $P4_2$ there are two possible equally valid ways of indexing the frames, related by e.g.

```
h' = k,  k' = h,  l' = -l
```

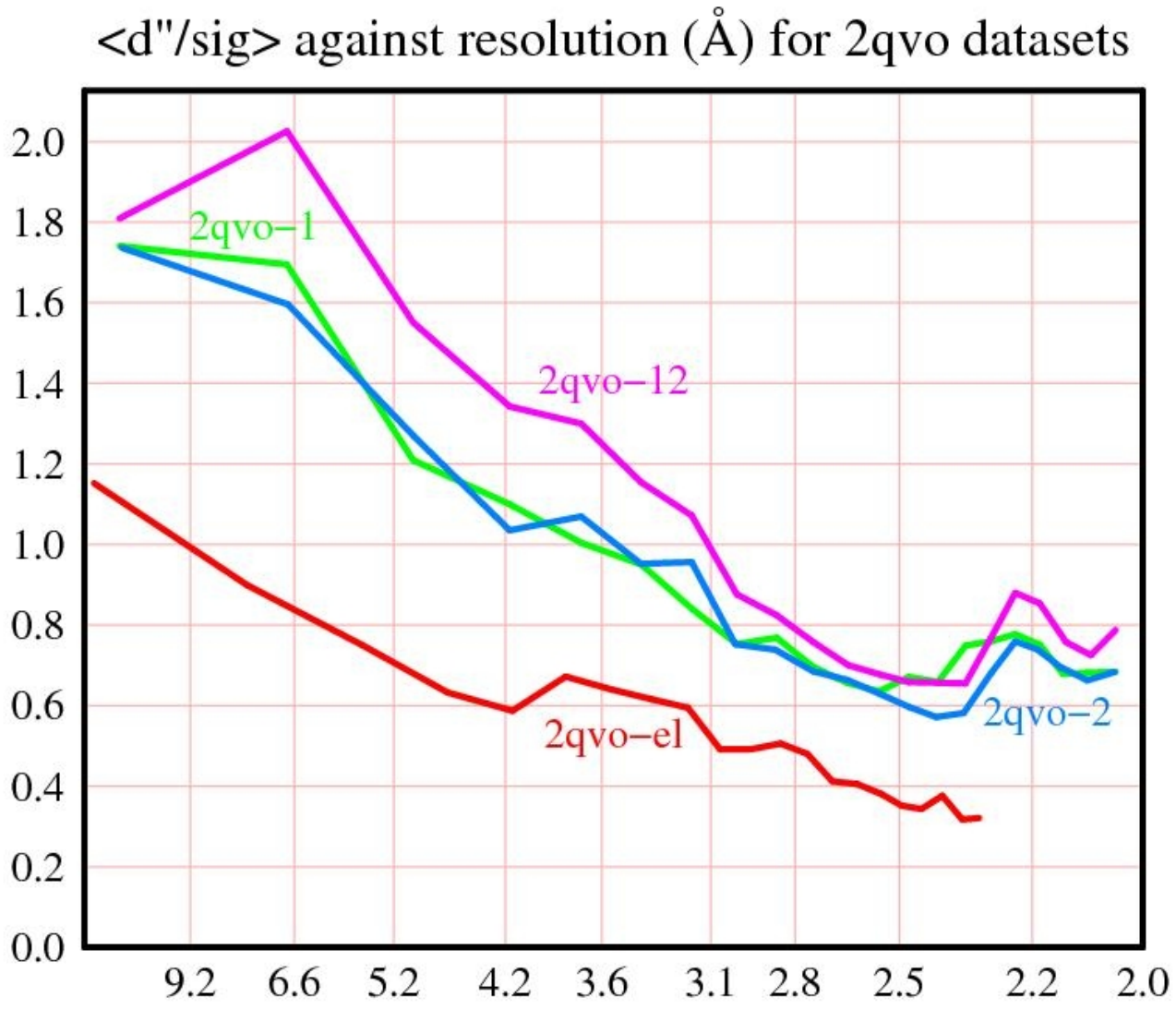
For a first-time structure solution using a single SAD dataset it does not matter which indexing is used, however it does matter if more than one dataset are combined, e.g. a long wavelength dataset high redundancy modest resolution dataset for better sulfur-SAD phasing and a low redundancy higher resolution native dataset for better density modification and structure refinement. In this case all three datasets provided happen to be indexed the same way but it is the alternative to the desposited PDB file 2qvo. To make it easier to check the quality of the results, the deposited unit-cell was used throughout as in the XDS processing (a=b=53.03, c=40.97 Å), and the data were reindexed to the deposited setting.

The data provided were first converted from mtz to sca format using Tim Gruene's MTZ2SCA:

```
mtz2sca xds-2qvo-1-1_360-I.mtz
```

(in the case of the XDS data for 2qvo-1) and the resulting .sca file reindexed with XPREP to give 2qvo-1.sca, and similarly for 2qvo-2.sca (XDS), 2qvo-12.sca (XDS, both runs merged) and 2qvo-el.sca (ELVES).

The best way of assessing the quality of one individual SAD dataset, given high redundancy, is probably a plot of the CC (correlation coefficient) between two randomly chosen subsets of the anomalous differences. This has the advantage that the esds of the intensities, which are difficult to estimate accurately, are not used. Unfortunately the data provided for this workshop have already had equivalent reflections, except for Friedel opposites, merged so we cannot use this criterion. Another approach is to plot the mean value of $\{|I_+-I_-|/esd(I_+-I_-)\}$, where '+' and '-' refer to hkl and -h-k-l respectively, against the resolution. If there is no significant anomalous signal for the highest resolution data, this should asymptote to the statistical value of $\sqrt{(2/\pi)} = 0.798$ (Bunkoczi & Debreczeni, personal communication). Since weak SAD data are noise at the high resolution end, we can use this to check the scale of the intensity esds. These plots also enable us to decide on a possible resolution cutoff at which to truncate the data to avoid letting in too much noise. The results for the four 2qvo datasets are shown in the Figure. However for 'low hanging fruit' like this structure, the SHELXC and HKL2MAP default value of maximum resolution plus 0.5 Å works surprisingly well, so we will use that value here. The diagram was produced using XPREP (version 2011/1), but HKL2MAP and other programs can produce similar plots, and the necessary numerical data are also output by SHELXC.



The ELVES data do not extend to such a high resolution as the XDS datasets, but the extra XDS data probably do not contain much anomalous signal. None of the datasets have optimal error models, though ELVES (red) is clearly underestimating the esds of the reflection intensities. It is not obvious from this plot that 2qvo-2 is better than 2qvo-1, but we shall see that it is.

Normally SHELXC/D/E are run using HKL2MAP or other GUIs, but for educational reasons we will run them from a command line here. This requires setting up a small script using a text editor, rather like the scripts that we used to run CCP4 in a previous millenium. For the 2qvo data we will call this file 2qvo-1. Note that it calls the multprocessor version of SHELXD (beta-test available by email request). The normal version would also work fine but is slower (even if there is only one CPU). Here is the script:

```
shelxc 2qvo-1 <<EOF
cell 53.03 53.03 40.97 90 90 90
spag P42
sad 2qvo-1.sca
sfac S
find 4
mind -3 2
ntry 1000
EOF
shelxd_mp 2qvo-1_fa
```

The scripts 2qvo-2, 2qvo-12 and 2qvo-el are made by replacing 2qvo-1 with 2qvo-2 throughout etc. Note that the cell and space group given here are used in preference to those in the .sca file. The MIND instruction eliminates heavy atoms closer than 3 Å from stronger peaks or closer than 2 Å from a symmetry equivalent of the same atom. The script is made executable and started as follows (Linux/Mac):

```
chmod ugo+x 2qvo-1
./2qvo-1
```

SHELXC writes three files in addition to some useful statistics that are output to the console. These files are:

2qvo-1.hkl – merged native intensity data for density modification with SHELXE (and possibly later refinement with SHELXL). Friedel opposites have also been merged.

2qvo-1_fa.hkl – reflection indices, estimated F_A values and phase shifts α. For SAD experiments, F_A is approximated to by $|F_+ - F_-|$ and α by 90º when $|F_+| > |F_-|$ and 270º when $|F_-| > |F_+|$. This file is read by both SHELXD and SHELXE. When the heavy atoms have been found and their substructure phases ϕ_H calculated, the starting native protein phases ϕ_P can be estimated in SHELXE by:

$$\phi_P = \phi_H + \alpha$$

these starting phases are then improved by density modification (using the *sphere of influence algorithm*) and (in the new beta-test version) by iterative *poly-Ala chain tracing*.

2qvo-1_fa.ins – the instruction file for SHELXD. It contains:

```
TITL 2qvo-1_fa.ins SAD in P42
CELL  0.98000  53.03  53.03  40.97  90.0  90.0  90.0
LATT  -1
SYMM -Y, X, 1/2+Z
SYMM -X, -Y, Z
SYMM Y, -X, 1/2+Z
SFAC S
UNIT    64
SHEL 999 2.5
PATS
FIND 4
MIND -3 2
NTRY 1000
SEED 1
HKLF 3
END
```

We have asked shelxd to find 4 heavy atoms: 3 Met sulfurs and one Cys sulfur. In fact there are five sites because one of the methionines (Met91) is disordered, but SHELXD is not so easily fooled. PATS sets up Patterson seeding and NTRY the number of trials. With the number of trials it is wise to play safe and the multiprocessor version of SHELXD is *very* fast. These SHELXD runs for all four datasets gave the five correct sites as the top five in the list, for example file 2qvo-2_fa.res:

```
REM Best SHELXD solution:   CC 42.07   CC(weak) 26.76   CFOM  68.83
REM
TITL 2qvo-2_fa.ins SAD in P42
CELL  0.98000   53.03   53.03   40.97   90.00   90.00   90.00
LATT  -1
SYMM -Y, X, 1/2+Z
SYMM -X, -Y, Z
SYMM Y, -X, 1/2+Z
SFAC S
UNIT    64
S001   1  0.661392  0.969666  0.218945   1.0000  0.2
S002   1  0.624695  1.153847  0.273583   0.8739  0.2
S003   1  0.836647  1.218018  0.355276   0.5519  0.2
S004   1  0.518761  1.133705  0.259364   0.5018  0.2
S005   1  0.562004  1.182877  0.268695   0.4483  0.2
S006   1  0.816055  1.156021  0.220466   0.2864  0.2
HKLF 3
END
```

Although not required for structure solution, a MOVE instruction was inserted before the first site in each .res file so that all solutions would have the same hand and unit-cell origin as the deposited PDB

file. In this case

```
MOVE 0 0 0.21779 -1
```

was required, corresponding to the transformation **-x, -y, 0.21779-z**. In case this gets overwritten when the skript is rerun, backup copies are provied of these .res files (bck-*). As a result of the MOVE instruction, the original PDB entry 2qvo.ent can be displayed directly on top of the native density (from *.phs) or anomalous density (from *.pha) using e.g. COOT.

The command for density modification with iterative poly-Ala tracing using the SHELXE beta-test (available free on email request for registered SHELX users) was

```
shelxe 2qvo-2 2qvo-2_fa -s0.54 -h -q -a -e1.4
```

and similarly for the other three datasets. Normally the inverse heavy atom substructure would also have to be tested with **-i**, but the MOVE command ensures that the phasing starts from the correct heavy atom enantiomorph in this case. SHELXE reads the files **2qvo-2.hkl** (native reflection data), **2qvo-2_fa.hkl** (hkl, F_A and α) and **2qvo-2_fa.res** (see above). The command-line switches have the following meanings:

-s0.54 – solvent content. In this case taken from the deposited PDB file, but assuming that each amino-acid occupies 140 $Å^3$ gives an acceptable result (Kevin Cowtan, personal communication).

-h5 – use the first five heavy atoms. **-h** would have used all.

-q – search for alpha-helices as well as general tripeptides. If one is sure that no alpha-helices are present, this may be left out to save time.

-a – three iterations of poly-Ala tracing. Otherwise the number has to be specifies, e.g. **-a5**.

-e1.4 – extend the data to 1.4 Å in the last iteration using the *free lunch algorithm*. Often **-e1** is used, but here a larger value was used to avoid creating very large .phs (native phases) and .pha (anomalous phases) files for downloading. This usually improves the map quality, sometimes dramatically, but works best when the native data extent to 2.0 Å or better.

SHELXE wrote the following files in this example:

2qvo-2.lst – listing file.

2qvo-2.phs – final native amplitdes and phases.

2qvo-2.pha – final substructure amplitudes and phases.

2qvo-2.pdb – poly-Ala trace.

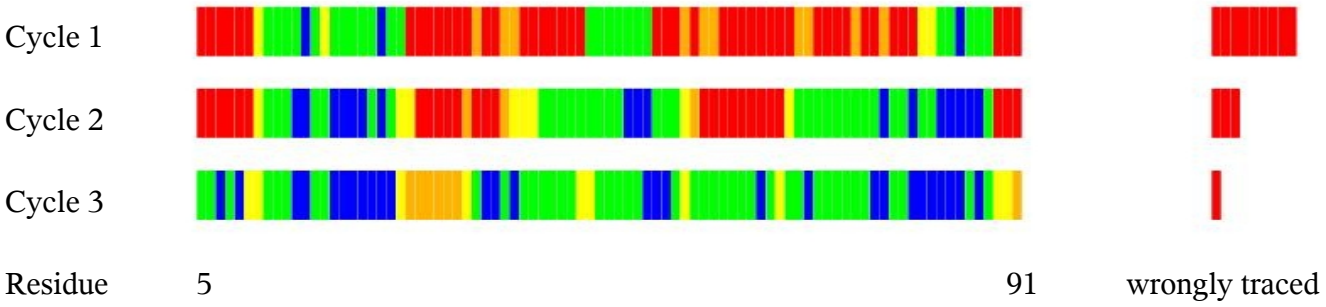
These three files are suitable for displaying the results with COOT. In this case the anomalous phases show the disorder of the sulfur atom of Met91 particularly clearly.

2qvo-2.hat – revised heavy atoms. In difficult cases this can be renamed as the *_fa.res file and the SHELXE job repeated, but that is not necessary here. In general the recycled heavy atoms are a little more precise than those from shelxd because they indirectly use information from the native data as well as the anomalous differences, and this is also a way to add additional weak sites. However recycling can be counterproductive if the DSUL instruction is used in SHELXD to resolve disulfide units. Note that in the latest SHELXE, **-b** is no longer required to produce this file.

The various correlation coefficients for the four datasets were:

Dataset	**2qvo-el**	**2qvo-1**	**2qvo-2**	**2qvo-12**
shelxd CC	36.5	35.7	42.1	43.8
ahelxd CC(weak)	24.0	19.8	26.8	25.3
CFOM = CC+CC(weak)	60.5	55.5	68.8	69.0
Poly-Ala trace cylce 1	19.9	19.5	30.2	39.6
Poly-Ala trace cycle 2	35.6	38.5	48.0	48.6
Poly-Ala trace cycle 3	48.1	44.8	48.3	47.9
#residues/#chains	87/2	88/2	82/1	87/4

Note that the first block of CC values is based on the anomalous differences, the second on the native structure factors. Although the two weaker datasets start with lower values, after (the default) three autotracing cycles all give very satisfactory traces and maps. Reliable criteria for a correct solution are that the mean chain length should be greater than 10 amino-acids and the CC against the native data greater than 25. There are 95 amino-acids in the sequence but the PDB entry 2qvo contains only 87. The following diagram illustrated the model improvement during the poly-Ala recycling (for 2qvo-el) as a function of residue number in the sequence:



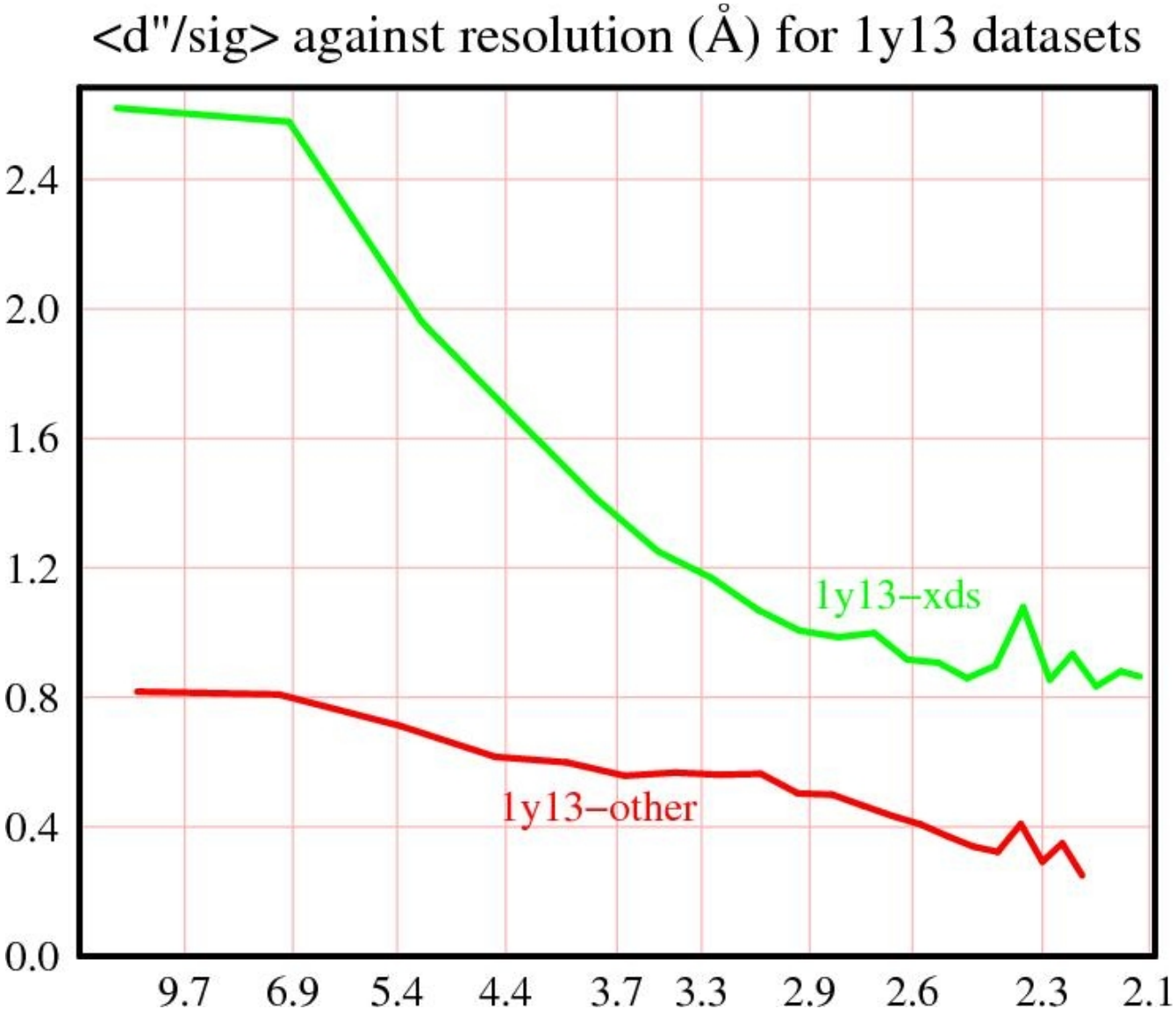
The color code for individual residues is as follows: blue: Cα within 0.3 Å, green within 0.6 Å, yellow with 1.0 Å, orange within 2.0 Å and red not traced. The number of residues incorrectly fitted (Cα > 2Å) is shown in red on the right on the same scale.

1y13 weak anomalous test

These data have problems in the correct definition of the experiments performed and also suffer from severe radiation damage. A processed dataset of unknown origin was also provided (we will call it 1y13-other.sca) and there is a very nice detective report:on the raw data by Kay Diederichs at

http://strucbio.biologie.uni-konstanz.de/xdswiki/index.php/1Y13

Kay was able to sort out the problems with the data using XDS and also showed how HKL2MAP and SHELXC/D/E could be used to obtain a relatively complete trace. We use XPREP as in the previous example to assess the anomalous signal in both processed datasets.



The XDS data asimptote well to the theoretical value of $\sqrt{(2/\pi)} = 0.798$ and show substantial anomalous signal at lower resolution. Something is wrong with the estimates of the reflections esds in the 'other' dataset and it shows only weak anomaous signal. Preliminary attempts to solve the structure with the 1y13-other data were not successful and there seems little point in wasting time on data that are known

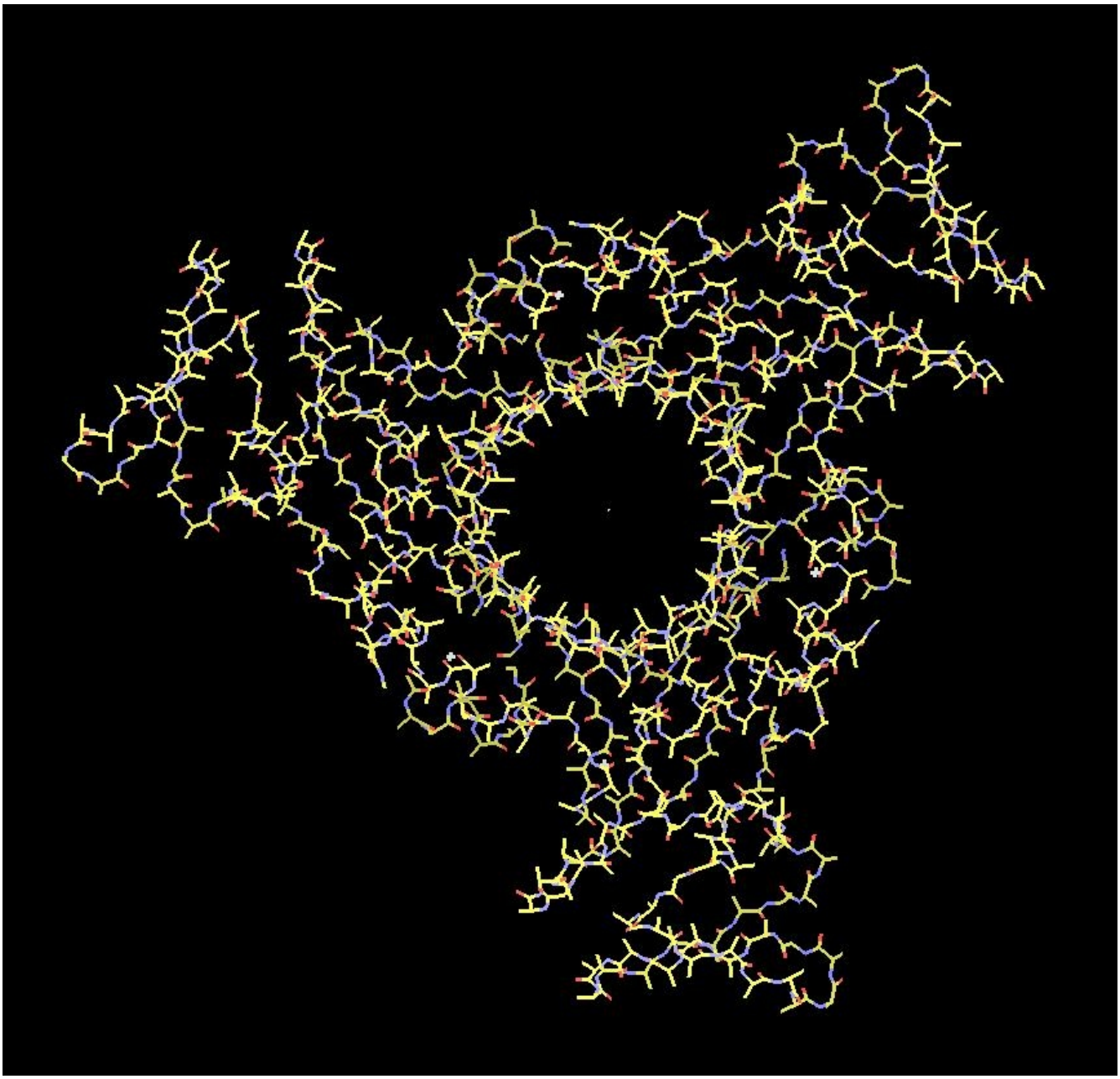
to contain serious experimental errors that could have been corrected. We will use the XDS data exclusively from now on. The following script turns out to be more or less optimal for the complete phasing:

```
mtz2sca xds-1y13-raddam-I.mtz
shelxc 1y13 <<EOF
cell 103.319 103.319 131.111 90 90 90
spag P43212
sad xds-1y13-raddam-I.sca
find 6
sfac Se
ntry 1000
mind -3 2
EOF
shelxd_mp 1y13_fa
shelxe 1y13 1y13_fa -s0.58 -m50 -q -h6 -a -n3 -e1 -l4
```

The multi-CPU version of shelxd was used to save time. It produced the following file 1y13_fa.res. Cutting back the rsolution with `SHEL` would have given larger CC values but the default truncation at maximum resolution + 0.5 Å gave the expected six clear atoms and probably more precise coordinates; this explains why recycling this heavy atom solution later did not improve the results.

```
REM Best SHELXD solution:   CC 27.79   CC(weak) 19.51   CFOM  47.30
REM
TITL 1y13_fa.ins SAD in P43212
CELL  0.98000  103.32  103.32  131.11   90.00   90.00   90.00
LATT  -1
SYMM 1/2-Y, 1/2+X, 3/4+Z
SYMM -X, -Y, 1/2+Z
SYMM 1/2+Y, 1/2-X, 1/4+Z
SYMM 1/2-X, 1/2+Y, 3/4-Z
SYMM Y, X, -Z
SYMM 1/2+X, 1/2-Y, 1/4-Z
SYMM -Y, -X, 1/2-Z
SFAC Se
UNIT   192
SE01   1  0.226768  0.755692  0.117831   1.0000  0.2
SE02   1  0.366028  0.844673  0.194222   0.8018  0.2
SE03   1  0.305832  0.454971  0.133449   0.7835  0.2
SE04   1  0.465775  0.823059  0.282689   0.7779  0.2
SE05   1  0.025383  0.825813  0.140069   0.7064  0.2
SE06   1  0.040421  0.976151  0.049447   0.5002  0.2
SE07   1  0.247734  0.780983  0.122898   0.2653  0.2
SE08   1  0.251472  1.037659  0.188357   0.1454  0.2
HKLF 3
END
```

In fact three of the heavy atoms are zinc and three are selenium, but for SHELXE it does not matter which are which, to a first approximation the occupancies allow for the different scattering powers. SHELXE – perhaps with a little luck – after 2 poly-Ala tracing cyles has found three chains of 157, 155 and 160 residues compared with three chains of 163 residues in the deposited PDB file. Since the space group is known to be $P4_32_12$ rather than $P4_12_12$ inversion of the heavy-atom structure is not required. For the free lunch expansion to 1.0 Å it was necessary to assign more memory with **`-l4`**. Although the exploitation of NCS in SHELXE is rather quick and dirty, **`-n3`** finds about 50 extra amino-acids compared to non-NCS trials. The results could be fed into BUCCANEER or wARP to add the side-chains automatically, but it is also easy and more educational to do this with COOT. A COOT screenshot of the SHELXE ploy-Ala trace is shown below, the three-fold NCS is clearly apparent.



Eviltwins

Three artificially constructed datasets were provided for a pseudomerohedral twinning test. It turns out that the 60:40% data indeed corresponded to the sum of two differently orientated diffraction patterns in this ratio corresponding to the PDB entry 1G1C. The cell and orientation of the data in the mtz file were consistent with this. This pseudo-merohedral twinning is possible because the **b** and **c** axes of 1G1C have almost the same length. Tim Gruene's MTZ2SCA was used to prepare the file 6040.sca. Despite the (artificial) twinning, XPREP had no problem in finding the same space group as 1G1C ($P2_12_12_1$) using the 60:40% data.. However the datasets lebeled as 50:50% and 80:20% turn out to be identical and correspond to a 50:50% twin. The given cell was different to that for the 60:40% data and, unlike the 60:40% data, these data have to be reindexed using the matrix: 0 0 1; 1 0 0; 0 1 0 to bring them into the same orientation as 1G1C.

In fact XPREP is fooled by the twinning for the 50:50% data. The good tetragonal statistics (R_{merge} = 3.1%) led it to suggest the space group $P4_32_12$ for the 50:50% data. The $P4_32_12$ systematic absences that might have prevented this had apparently been removed in the data processing, never a good idea! XPREP may even have suspected that the data were tetragonal lysozyme that has almost identical cell dimensions and is really $P4_32_12$.

However there is one number in the XPREP output for both datasets that should have set off all the alarm bells: ***the mean value of $|E^2-1|$ is 0.551 for the 60:40% data and 0.536 for the 50:50% data***, both much lower than the expected value of 0.736. This number can sometimes be too high, e.g. in the presence of translational NCS, but never too low. Such a low value has only two common explanations: (a) F-values have been read in but treated as intensities, or (b) the data are twinned.

A routine attempt was made in both cases to locate the heavy atoms (four methionine sulfurs) using the anomalous differences.

```
shelxc 6040 <<EOF
cell 38.299 79.088 79.107 90 90 90
spag P212121
sfac S
find 4
sad 6040.sca
mind -3 2
ntry 1000
EOF
shelxd_mp 6040_fa
```

Surprisingly, this finds the correct four sulfur atoms as the four strongest peaks! However attempts to extend this to the full native structure were not successful, clearly this problem is not a low hanging fruit and more sophisticated programs will be needed. In the case of the 50:50% twin, a similar attempt did not even find the sulfur atoms.

SHELXL refinements against the evil twinned data

SHELXL was originally written for small molecules and is only really suitable for the refinement of macromolecules that diffract to about 1.6 Å or better. Twinning reduces the effective data to parameter ratio so an even higher resolution is needed for refinement of a twin. For the 50:50% twin, the data tto praameter ratio is halved so the data behave as if they had a maximum resolution of $1.93*2^{(1/3)} = 2.43$ Å, rather than the 1.93 Å from the data processing. This is well outside the acceptable range for SHELXL refinements. SHELXL lacks in particular the trosion angle restraints necessary for side-chain refinement, but also the solvent model is inadequate for low resolution refinement.

Even though there are better programs for the purpose, we will use these data to illustrate how to set up a SHELXL refinement. SHELXL needs only two input files. In fact we have just created a .hkl merged native data files in the right format using SHELXC, but we need to flag the free-R reflections set, e.g. using XPREP. In future the .hkl files from SHELXC will also contain free-R flags. Then we need to run SHELXPRO in order to read a PDB format file (which uses orthogonal coordinates) and write a SHELXL .ins file (based on crystal coordinates). The PDB file for 1g1c has been downloaded as 1g1c.ent, but we will generate `6040-1.ins` (and `6040-1.hkl`) for the refinement. Unless otherwise shown (in italics) the questions are simply answered with <Enter>.

`shelxpro 6040-1`

```
SHELXPRO - SHELX interface for protein applications - Version 97-3
Copyright(C) George M. Sheldrick 1996-2003

[F] New output filename                    [V] R(free) files
[A] Anisotropic scaling (Hope & Parkin)    [I] .ins from PDB file
[P] Progress of LS refinement diagram      [L] Luzzati plot
[T] Thermal displacement analysis          [E] Esd analysis
[U] Update .res (and .pdb) to .ins file    [N] NCS analysis
[R] Ramachandran Phi-Psi plot              [K] Kleywegt NCS plot
[M] Map file for O from .fcf               [O] PDB file for O
[H] .hkl file from other data formats      [Y] X-PLOR/CNS .fob to .hkl
[D] Convert DENZO/SCALEPACK .sca to .hkl   [C] Color plots (now on)
[X] Write XtalView map coefficients        [W] Write Turbo-Frodo map
[S] Reflection statistics from .fcf        [Z] Least-squares fit
[J] Generate restraints from model         [B] PDB deposition
[G] Generate PDB file from .res or .pdb    [Q] Quit

Enter option: I

Reads a PDB file and generates a SHELXL .ins file.  The PDB file is assumed
to conform strictly to the PDB format as defined by the Protein Data Bank,
but closely related non-standard formats (e.g. CCP4 and XPLOR) can usually
be understood.  The program will ask for the missing cell and symmetry
information etc.  Engh and Huber restraints are included in the .ins file
for standard residues, and extra restraints are added for disulfide bridges
and C-terminal carboxyl groups.  A summary of the residue and atom names is
written to the .pro file for subsequent reference.

** The I option is intended for initial input of a structure to SHELXL,
NOT for updating between refinement jobs, for which 'U' should be used. **

Enter N to abort option, <Enter> to continue:
```

```
Enter name of .ins file [6040-1.ins]:
Enter name of PDB file [6040-1.ent]: 1g1c.ent
Enter title [6040-1]:
CELL in Angstroms and deg. [38.300 78.600 79.600 90.00 90.00 90.00]:

Enter Z (number of molecules per cell) [8]:
Enter space group in PDB or XPREP notation [P 21 21 21]:
Enter wavelength in Angstroms [1.54178]:
Generate atom coordinates using SCALE instructions from PDB file (P) or use
current cell to calculate transformation matrix (C) [C]:

SHELXL does not recognize chain ID letters, so it will be necessary to
incorporate these into the residue numbers by adding offsets of (say) 1000
for chain A, 2000 for B etc. to the residue numbers from the PDB file.
Offset for chain A [1000]:
Offset for chain B [2000]:

Enter old residue numbers (modified by chain ID, if any) for all N-terminii
(<CR> if none). To continue on the next line, put "=" at the end of the line
: 1002 2002

Enter old residue numbers for all C-terminii in the same way: 1099 2099

Enter old residue numbers in the same way at which renumbering of a block of
residues should start. The block continues until the next residue specified
here (<CR> if none):

New residue number for first solvent water [3001]:

Reset water occupancies to unity (Y or N) ? [Y]:

HKLF code (3 for F, 4 for F-squared) [4]:

The .ins file has been written successfully.  The U option in SHELXPRO may
be used for further checking of occupancies etc.
```

Note that although residue 99 is a true C-terminus, 2 is not the N-terminus, but residue 1 was not visible in the density. All that specifying 2 as the N-terminus will cause is that a HFIX instruction will be included to turn it into $-NH_3^+$ later. We have now set up a valid SHELXL refinement (mainly by hitting <Enter>), but in this case it is a good idea to edit the following lines. We are lucky that no special ligands have to be defined, but the PRODRG server could have been used to generate the necessary SHELXL format restraints in such a case. The recommended edits (SHEL and SIMU) and insertions (TWIN and BASF) are:

```
SHEL 10 0.1  →  SHEL 99 0.1    ! no low resolution citofff needed

SIMU 0.1 $C_* $N_* $O_* $S_*
  → SIMU 0.02 $C_* $N_* $O_* $S_*  ! tighten restraint for low resolution

TWIN  1 0 0  0 0 1  0 -1 0 ! refine as twin
BASF 0.1                   ! starting value for twin factor (should actually be 0.4)
```

The refinement is started by:

SHELXL 6040-1

Note that in this case we have not used the multi-CPU version (developed by Kay Diederichs) although it is much faster, because this sometimes has problems with the refinement of twins. These refinements gave the following results:

```
6040 R1 = 15.3% for F>4sigma(F) and 18.4% for all data;
     R1free = 22.4% for F>4sigma)F) and 24.7% for all data
BASF (twin factor) = 0.426

5050 R1 = 15.1% for F>4sigma(F) and 17.8% for all data;
     R1free = 22.0% for F>4sigma(F) and 24.4% for all data
BASF (twin factor) = 0.496
```

For reasons explained above, the gap between R and Rfree is too large (and becomes slightly larger if more refinment cycles are performed. Otherwise the results are fine. They can be viewed using COOT by reading in the .res file followed by the .fcf file created by SHELXL. In theory one can perform some editing and output the .ins file for the next refnement job from COOT, but in practice COOT does not always write this file correctly and it is usually advisable to hand-edit it after writing it with COOT (please report problems to the COOT users' list, not me).

In this case the instruction files for consecutive jobs would be nubered 6040-1.ins, 6040-2.ins, etc. and it would also be necessary to create links to or copies of the corresponding .hkl files; these do not change during the refinement.