CCP4 Study Weekend 2017

An introduction to experimental phasing of macromolecules

January 10th 2017

George M. Sheldrick

http://shelx.uni-ac.gwdg.de/SHELX/
The crystallographic phase problem

Although the crystallographic phase problem would appear to be difficult to solve, the vast majority of small-molecule structures are solved routinely in a few seconds by black box direct methods. These assume that the structure consists of resolved atoms, which means that data to a resolution of 1.2Å or better is required.

Few new macromolecular structures diffract so well, but a recent success was the direct methods solution of poly-adenosine, a RNA that had defied other attempts at solution but diffracted to almost 1.0Å.
The other double helix

In 1961 Rich, Davies, Watson & Crick proposed a parallel double-helix structure for polyadenylic acid based on fibre diffraction photographs, using similar reasoning to that used for the antiparallel double helix of DNA:
The \((rA_{11})_2\) parallel double helix

Over 50 years later, the proposed structure turned out to be remarkably accurate. The bases are related by a non-crystallographic twofold rotation.


Experimental phasing of macromolecules

Except in the relatively rare cases where atomic resolution data permit the phase problem to be solved by *ab initio* direct methods, experimental phasing usually implies the presence of ‘marker’ (heavy) atoms. These may be present naturally (e.g. metals or sulfur) or may have added (e.g. iodide soaks or selenomethionine derivatives).

In order to locate the marker atoms, we must first estimate the substructure structure factors $F_A$. In the case of SAD phasing, we can use the anomalous differences as a crude approximation, and then use a small molecule direct methods program to find the marker atoms.

The first to use direct methods to find the heavy atoms from isomorphous differences was Thomas Steitz (*Acta Cryst.* (1968) **B24** 504-507), but it was Keith Wilson who made the approach accessible to ordinary mortals by applying the program MULTAN, at that time the direct methods program of choice for small molecules (*Acta Cryst.* (1978) **B34** 1599-1608).
SHELX direct methods programs

SHELXD was originally written to solve large chemical structures at atomic resolution (d < 1.2Å). The real space part of the dual space cycle imposes a strong atomicity constraint on the phases that are then refined using phase probability relations in the reciprocal space part.

By chance, SHELXD turned out to be rather effective at locating heavy atoms in macromolecular structures. The lower resolution does not matter because the heavy atoms are still resolved from one another. For small molecules, SHELXT gives a more complete solution and also determines the space group, but SHELXD lives on for macromolecules, e.g. in several CCP4i2 pipelines.
The analysis of MAD data

Karle (1980) and Hendrickson, Smith & Sheriff (1985) showed by algebra that the measured intensities in a MAD experiment, assuming only one type of heavy atom, should be given by:

\[ |F_\pm|^2 = |F_T|^2 + a|F_A|^2 + b|F_T||F_A|\cos \alpha \pm c|F_T||F_A|\sin \alpha \]

where \( a = (f''^2 + f'^2)/f_0^2 \), \( b = 2f'/f_0 \), \( c = 2f''/f_0 \) and \( \alpha = \phi_T - \phi_A \). This needs \( f' \) and \( f'' \) for each wavelength, though in practice they go out in the wash.

In a 2-wavelength MAD experiment, we have 4 equations for the 3 unknowns \( F_A, F_T \) and \( \alpha \), so with error-free data we should get a perfect map! For SIRAS, we have 2 equations for the derivative plus one for the native data. Given perfect isomorphism, the phase problem is also solved.

**SAD phasing involves extracting these three numbers from two observations, at first sight an impossible task.**
Why does SAD phasing work?

We would like to be able to use $F_A$, the heavy atom structure factor, to solve the substructure, but instead we have to make do with:

$$|F_+| - |F_-| = c |F_A| \sin \alpha$$

Which can be derived from the MAD equations by assuming that the anomalous differences $||F_+| - |F_-||$ are small. $||F_+| - |F_-||$ is normalized to get $E$-values. This eliminates $c$ and its resolution dependence! But we still have one data item for two unknowns, $|F_A|$ and $\alpha$. We need $|F_A|$ to find the marker atoms and $\alpha$ to get starting values for the native phases using $\phi_T = \phi_A + \alpha$.

The trick is to use only the largest $||F_+| - |F_-||$ and to assume that they correspond to $\alpha \approx 90^\circ$ (when $I_+ >> I_-$) and $270^\circ$ ($I_- >> I_+$). Direct methods only uses the larger $E$-values anyway. Amazingly, this often works!

At low resolution, the better phase estimates from MAD rather than SAD can still make a decisive difference.
Using the marker atoms to estimate phases

SHELXC/D/E simplify the phase problem by making several severe approximations, e.g. that there is only one type of generic anomalous scatterer. The phases $\phi_A$ calculated for the heavy atom substructure are used to estimate starting phases $\phi_T$ for the full macromolecular structure by:

$$\phi_T = \phi_A + \alpha$$

where $\alpha$ is estimated from the diffraction data as already explained. For SAD these phases are clearly inferior to those from MAD, because they can only be calculated when the anomalous differences are large and $\alpha$ can only be estimated as $90^\circ$ or $270^\circ$. For reflections in centrosymmetric projections, SAD phases have a twofold ambiguity.

After improving these phases by density modification we can apply them to the native structure factors $F_T$ to calculate an electron density map. For SAD and SIR density modification is essential, for MAD it is much less important.
Anomalous differences are usually truncated to expedite substructure solution. $|\Delta F|/\sigma(\Delta F)$ requires good estimates of the intensity esds, but if it extrapolates at high resolution to the pure noise value of $\sqrt{(2/\pi)} = 0.798$, the esds must be OK. CC between two random $\Delta F$ subsets (CC1/2) is independent of the esds, but requires unmerged data.
Starting atoms consistent with Patterson

**Dual space recycling for SHELX XD substructure solution**


### SF calculation

**reciprocal space:**
refine phases

**real space:**
select atoms

Many cycles $E > E_{\text{min}}$

The $F_A$ values are normalized to $E$-values. The correlation coefficient $CC$ between $E_{\text{obs}}$ and $E_{\text{calc}}$ is calculated for all $E$-values. $CC_{\text{weak}}$ uses only the data with $E_{\text{min}} > E$ that were not used for substructure solution; the program chooses the solution with the highest $CFOM = CC + CC_{\text{weak}}$
The heavy atom enantiomorph problem

The location of the heavy atoms from the $|F_A|$-values does not define the enantiomorph of the heavy-atom substructure; there is exactly a 50% chance of getting the enantiomorph right. When the protein phases are calculated from the heavy atom reference phases, only one of the two possible maps should look like a protein, in favorable cases this enables the correct heavy atom enantiomorph to be chosen.

If the space group is one of an enantiomorphic pair (e.g. P4$_1$2$_1$2 and P4$_3$2$_1$2) the space group must be inverted as well as the atom coordinates.

For three of the 65 Sohnke space groups that are possible for chiral substructures, the coordinates have to be inverted in a point other than the origin! These space groups and inversion operations are:

\[ I4_1 (1-x, \frac{1}{2}-y, 1-z); \quad I4_{122} (1-x, \frac{1}{2}-y, \frac{1}{4}-z); \quad F4_{32} (\frac{1}{4}-x, \frac{1}{4}-y, \frac{1}{4}-z). \]
SHELXD histograms and occupancies for Elastase

1 solution in 100 trials

Tendamistat CC, CCweak and hits per 10000 tries

Occupancy (from SHELXD) of super-sulfur peak 8

Occupancy of peak 9
Disulfide bond resolution

When the anomalous signal does not extend to sufficient resolution to resolve disulfides, it has been standard practice to search for super-sulfur atoms.

An effective alternative (DSUL) is to modify the peaksearch to locate the best positions for S-S units in the slightly elongated electron density maxima. These resolved disulfides not only improve the performance of the substructure solution, they also give a much better phase extension to higher resolution and better final map correlation coefficients. The extra CPU time required is negligible.
Critical parameters for SHELXD

1. It may be necessary to try different resolutions for data truncation.

2. The number of sites requested should be within about 20%.

3. If there is pseudo-translational NCS it may be advisable to switch off Patterson seeding (by leaving out PATS).

4. In the case of a soak, the rejection of sites on special positions should be switched off.

5. For S-SAD, DSUL (search for disulfides) is recommended for the resolution range 2.0 to 3.0Å. At lower resolution one should search for ‘super-sulfurs’ (disulfides).

6. In difficult cases it may be necessary to run more trials (say 50000).

SHELXD is highly parallel and scales up well on a multiple-CPU machine.
Fine tuning marker atom location

The procedure described here is very simple and robust, but achieves this with the help of drastic assumptions. In borderline cases it may be worth using the LLG (log likelihood gain) to refine substructure solutions, e.g. using the programs SHARP, CRANK2 or PHASER, before proceeding with density modification and tracing. In general LLG-based methods require more detailed information (e.g. which elements are present) than SHELXC/D/E, and tend to be slower. For details see:


Simulations with one heavy atom in P1

A perfect MAD or SIRAS experiment should give perfect phases! A centrosymmetric array of heavy atoms is fatal for SIR, but one heavy atom is enough for SAD even in space group P1, because it is easy to remove the negative image by setting negative density to zero!
Density modification

Although a high quality MAD dataset can produce an immediately recognizable map, for SAD it is essential to improve the map by density modification, which works best at high solvent content and high resolution.

Clearly, if we simply do an inverse Fourier transform of the unmodified density we get back the phases we put in. So we try to make a chemically sensible modification to the density before doing the inverse FFT in the hope that this will lead to improved estimates for the phases.

Classical density modification, pioneered by BC Wang, divides the map into protein and solvent regions, and flattens the solvent density. SHELXE uses instead the sphere of influence algorithm. If there is a large variation V in the density on a spherical surface 2.42Å (a typical 1,3-distance) from a voxel, that voxel is more likely to be an atomic position, and its density is retained. If the density variation on the sphere is small, the density of the voxel is ‘flipped’ (multiplied by -1).
The sphere of influence algorithm

The variance $V$ of the density on a spherical surface of radius 2.42 Å is calculated for each voxel in the map. The use of a spherical surface rather than a spherical volume was intended to save time and add a little chemical information (2.42 Å is a typical 1,3-distance in proteins and DNA). $V$ gives an indication of the **probability** that a voxel corresponds to a true atomic position.

Voxels with low $V$ are *flipped* ($\rho_s' = -\gamma \rho$ where $\gamma$ is usually set to 1.0).

For voxels with high $V$, $\rho$ is replaced by $[\rho^4/(\nu^2\sigma^2(\rho)+\rho^2)]^{\frac{1}{2}}$ (with $\nu$ usually 0.5) if positive and by zero if negative. This has a similar effect to the procedure used in the CCP4 program ACORN.

For *intermediate values of $V$*, a suitably weighted mixture of the two treatments is used.

An empirical weighting scheme for phase recombination is used to combat model bias.
Enantiomorph discrimination

The *contrast* is defined as the variance of $V$ (the variance of the density on a spherical surface of radius 2.42Å from a given voxel) calculated over all voxels. If there are regions of low $V$ (solvent) and high $V$ (protein) the contrast is high. This is a good test to see which heavy atom enantiomorph is correct (up to this point, density modification has to be applied to both). If the contrast differs strongly for the two hands, the structure has been solved!
The SHELXE autotracing algorithm

The poly-Ala chain tracing in SHELXE (*Acta Cryst.* D66 (2010) 479-485) is primarily designed for iterative phase improvement starting from very poor phases. The tracing proceeds as follows:

1. Find potential α-helices in the density and try to extend them at both ends. Then find other potential tripeptides and try to extend them at both ends in the same way.

2. Tidy up and splice the traces as required, applying any necessary symmetry operations.

3. Use the traced residues to estimate phases and combine these with the initial phase information using sigma-A weights, then restart density modification. The refinement of one B-value per residue provides a further opportunity to suppress wrongly traced residues.
Extending chains at both ends

The chain extension algorithm looks two residues ahead of the residue currently being added, and employs a simplex algorithm to find a best fit to the density at the atom centers as well at ‘holes’ in the chain. The quality of each completed trace is then assessed independently before accepting it.

Important features of the algorithm are the generation of a no-go map that defines regions that should not be traced into, e.g. because they are close to rotation axes or heavy atoms, and the efficient use of crystallographic symmetry. The trace is not restricted to a predefined volume, and the splicing algorithm takes symmetry equivalents into account.
Criteria for accepting chains

The following criteria are combined into a single figure of merit for accepting traced chains:

1. The overall fit to the density should be good.

2. The chains must be long enough (in general at least 7 amino-acids); longer chains are given a higher weight.

3. There should not be too many Ramachandran outliers.

4. There should be a well defined secondary structure ($\phi/\psi$ pairs should tend to be similar for consecutive residues).

5. On average, there should be significant positive density 2.9 Å from N in the N→H direction (to a hydrogen bond acceptor):

\[
\text{C} \quad \text{N} \rightleftharpoons \text{H} \cdots \cdot \cdot \cdot \text{O}
\]

\[
\text{C}_\alpha
\]
Fibronectin autotracing test

This structure illustrates the ability of the autotracing to start from a noisy sulfur-SAD map. Recycling the partial (but rather accurate) traces leads to better phases and an almost complete structure.

In the first cycle, 41% was traced with $C_\alpha$ within 1.0Å, 33% within 0.5Å and 4% false. After 3 cycles the figures were 94%, 87% and 0%. 

<table>
<thead>
<tr>
<th>Cycle 1:</th>
<th>Cycle 2:</th>
<th>Cycle 3:</th>
</tr>
</thead>
</table>

Residue number →

$C_\alpha$ deviation:

\[
\begin{align*}
< 0.3\text{Å} & < 0.6\text{Å} < 1.0\text{Å} < 2.0\text{Å} < \\
\end{align*}
\]
In this case, assuming threefold non-crystallographic symmetry (−n3) found 50 residues more than without NCS.
Is the structure solved?

If the CC for the structure factors calculated for the trace against the native data is better than 25%, and the resolution of the native data is at least 2.5Å, it is extremely likely that the structure is solved. Another good indication is whether one can see side-chains:
ANODE

ANODE (included in the current CCP4 distribution) reads a PDB format file to calculate $\phi_T$ and a file from SHELXC containing $F_A$ and $\alpha$. The heavy atom substructure phases are then calculated using $\phi_A = \phi_T - \alpha$. A Fourier map calculated with phases $\phi_A$ and amplitudes $F_A$ then reveals the heavy atom substructure from a SAD, MAD or SIRAS etc. experiment.

'anode lyso' would read the PDB format file lyso.ent and the lyso_fa.hkl file from shelxc. Lysozyme SAD-phased by four I3C 'sticky triangles' gave the following averaged anomalous densities:

Averaged anomalous densities (sigma)

<table>
<thead>
<tr>
<th>Density</th>
<th>Atom</th>
</tr>
</thead>
<tbody>
<tr>
<td>23.207</td>
<td>I2_I3C</td>
</tr>
<tr>
<td>23.103</td>
<td>I1_I3C</td>
</tr>
<tr>
<td>21.607</td>
<td>I3_I3C</td>
</tr>
<tr>
<td>2.705</td>
<td>SD_MET</td>
</tr>
<tr>
<td>2.302</td>
<td>S_EPE</td>
</tr>
<tr>
<td>2.300</td>
<td>SG_CYS</td>
</tr>
<tr>
<td>0.539</td>
<td>C9_EPE</td>
</tr>
<tr>
<td>0.289</td>
<td>C4_I3C</td>
</tr>
</tbody>
</table>

(EPE is HEPES buffer)
ANODE (continued)

This table is followed by a list of the highest unique anomalous peaks and the nearest atoms to each. In addition, a .pha file is written for displaying the anomalous density in COOT. This approach always produces density with the same unit-cell origin as the original PDB file. Where alternative reflection indexing is possible it may be necessary to take it into account (with the switch –i).

This figure clearly shows disulfides and other sulfurs, even though the anomalous data were too weak to find them using SHELXD.

Please note that there are better and more sophisticated ways of doing this (e.g. LLG maps from SHARP or PHASER), but ANODE is fast and very easy to use.
Downloads and acknowledgements

SHELX is available (free for academic use) via the SHELX homepage shelx.uni-goettingen.de and is part of CCP4. The homepage contains extensive documentation and links to related software etc.

I am very grateful to everyone who helped me to develop and test SHELX, in particular Isabel Usón, Claudia Millán, Massimo Sammito and the rest of Isabel’s group in Barcelona, Andrea Thorn, Tim Grüne, Thomas Schneider, Kay Diederichs, Anna Lübben, Raj Pannu and Pavol Skubak.

References:

