In-house experimental phasing

Bruker Users’ Meeting, 21st September, 2010

George M. Sheldrick, Göttingen University

http://shelx.uni-ac.gwdg.de/SHELX/
Experimental phasing of macromolecules

Except in relatively rare cases where atomic resolution data permit the phase problem to be solved by \textit{ab initio} direct methods, experimental phasing usually implies the presence of heavy atoms to provide reference phases $\phi_A$. We then calculate the starting phases $\phi_T$ of the full native structure by:

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

Where $\phi_A$ is the calculated phase of the heavy atom substructure. As we will see, $\alpha$ can be estimated from the experimental data. The phase determination requires the following stages:

1. Locate the heavy atoms using estimates $F_A$ of the structure factors of the heavy atom substructure.
2. Possibly refine the heavy atom parameters, then use them to calculate the reference phases $\phi_A$.
3. The native structure factors $F_T$ and starting phases $\phi_T = \phi_A + \alpha$ give an electron density map that is then improved by density modification.
SAD phasing

For experimental phasing we require $F_A$ (the heavy atom structure factor), $F_T$ (the native structure factor, including heavy atoms if present in the native data) and $\alpha$ for each reflection. *In theory* this is enough to solve the phase problem!

For a MAD or SIRAS experiment we have at least three observations per reflection and can deduce $F_A$, $F_T$ and $\alpha$ from the data. For SAD phasing we unfortunately only have two observations per reflection. However we can assume $|F_T| = \frac{1}{2}(|F_+| + |F_-|)$ and derive:

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

where $c = 2f''/f_0$; $c$ is eliminated when we normalize to get $E_A$.

Amazingly, this is sufficient to find the heavy atoms and to use them to estimate the native phases $\phi_T$ for some reflections. The key is to work with the larger anomalous differences that will tend to be those with $\alpha$ close to 90° or 270°, so for these reflections we can estimate $|F_T|$, $|F_A|$ and $\alpha$!
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 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 CPU time overhead is negligible.

This suggests that searching for other small fragments in the real space part of the dual-space recycling may be a good way of extending substructure solution to noisier data, provided that it can be done efficiently. Examples could be Fe$_4$S$_4$-clusters, Ta$_6$Br$_{12}^{2+}$ and the I$_3$C ‘sticky triangle’ of three iodine atoms (Tobias Beck).
I3C - a sticky magic triangle

This additive is particularly suitable for long-wavelength SAD phasing. It contains three iodine atoms in the form of an equilateral triangle with a side of 6.0 Å, which makes it easy to recognize in a heavy-atom solution. It attaches itself to proteins mainly via hydrogen bonds.

How *not* to measure data quality

The quality of diffraction data is often estimated by calculating:

\[
R_{\text{sym}} = \frac{\sum_{\text{unique}(hkl)} \{ \sum_{\text{equivalents}} | I - \langle I \rangle | \}}{\sum_{\text{unique}(hkl)} \{ \sum_{\text{equivalents}} I \}}
\]

Where \(\langle I \rangle\) is the mean intensity of a group of equivalents.

\(R_{\text{sym}}\) is also called \(R_{\text{int}}\) (to emphasize that it is calculated with \(I\) not \(F\)), \(R_{\text{linear}}\) (in HKL2000) or \(R_{\text{merge}}\), though the latter sometimes means that data from different crystals were merged. If only one equivalent is measured for a given reflection, it is not included in the numerator or denominator.

\(R_{\text{sym}}\) is *NOT* a good data quality indicator because it increases as the number of equivalents being merged increases, despite the fact that the merged data are more accurate if more reflections are merged!
How to measure data quality

$R_{\text{sym}}$ can be corrected to make it independent of the redundancy:

$$R_{r.i.m.} = \frac{\sum_{\text{unique}(hkl)} \{ \frac{N}{N-1} \}^{\frac{1}{2}} \sum_{\text{equivalents}} | I-I> | \} \sum_{\text{unique}(hkl)} \{ \sum_{\text{equivalents}} I \}}$$

where $N$ is the number of equivalents for a given unique reflection. $R_{r.i.m.}$ (Weiss & Hilgenfeld, 1997) is also called $R_{\text{meas}}$ (Diederichs & Karplus, 1997). Whereas $R_{r.i.m.}$ indicates the precision of the unmerged data, $R_{p.i.m.}$ (Weiss, 2001) indicates the precision of the merged data:

$$R_{p.i.m.} = \frac{\sum_{\text{unique}(hkl)} \{ (N-1)^{-\frac{1}{2}} \sum_{\text{equivalents}} | I-I> | \} \sum_{\text{unique}(hkl)} \{ \sum_{\text{equivalents}} I \}}$$

The quality of the anomalous differences

A similar $R$-index can be used for the anomalous differences:

$$R_{\text{anom}} = \frac{\sum_{\text{unique}(hkl)} \{ \sum | \langle I_{hk\ell} \rangle - \langle I_{-h-k-\ell} \rangle | \} \} }{\sum_{\text{unique}(hkl)} 0.5 \{ \langle I_{hk\ell} \rangle + \langle I_{-h-k-\ell} \rangle \} }$$

The ratio $R_{\text{anom}} / R_{\text{p.i.m.}}$ has been proposed as a possible indicator of the suitability of the data for SAD phasing. $R_{\text{r.i.m.}}$, $R_{\text{p.i.m.}}$ and $R_{\text{anom}}$ can be calculated and displayed graphically by XPREP version 2011/1; note that no intermediate merging should be performed, otherwise the values will be wrong!

$R_{\text{anom}}$ can also be calculated using earlier versions of XPREP by first merging the data except for Friedel opposites (option ‘S’), then calculating $R_{\text{int}}$ with option ‘A’ and setting $R_{\text{anom}} = 2 R_{\text{int}}$.

Other useful indicators of the quality of the anomalous differences are $< |\Delta F| / \sigma(\Delta F) >$ and the correlation coefficient between the signed $\Delta F$ values of two crystals or the same crystal at two wavelengths, where $\Delta F$ is $(F_{hk\ell} - F_{-h-k-\ell})$. 
Tendamistat CC, CCweak and hits per 10000 tries

Occupancy (from SHELXD) of peak 8

R_{anom} = 0.0270
R_{p.i.m.} = 0.0122

Occupancy of peak 9
Density modification

The heavy atoms can be used to calculate reference phases; initial estimates of the protein phases can then be obtained by adding the phase shifts $\alpha$ to the heavy atom phases.

These phases are then improved by density modification. 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.

Many such density modifications have been tried, some of them very sophisticated. Major contributions have been made by Kevin Cowtan and Tom Terwilliger. One of the simplest ideas, truncating negative density to zero, is actually not too bad (it is the basic idea behind the program ACORN).

The program SHELXE uses a rather simple density modification algorithm to obtain a map quickly and robustly, rather than trying to get the best possible phases.
**SHELXE density modification**

The most interesting aspects of the SHELXE density modification are *what it doesn’t do!* SHELXE does NOT do or use the following:

**Solvent flattening.** In fact SHELXE doesn’t know where the solvent is!

**Masks** for the solvent or NCS regions (slow, error-prone and messy).

**Histogram matching.** This implies that one knows what to match the histogram to (e.g. the histogram of a ‘similar’ structure).

Since these have not been exploited directly, the presence of flat solvent regions in the resulting map or a good fit to an expected histogram are useful indications of the quality of a map.
Detecting PO$_4$ groups

To introduce the *sphere of influence* concept, we first look at the method used by Tim Gruene for the location of phosphate units in DNA and RNA structures. It can also be used to find phosphate and sulfate ions in the solvent region.

A *sphere of influence* is generated around the putative phosphorus position (an electron density peak) with a radius equal to a P–O bond length. The correlation coefficient CC is calculated between the electron density at each point on the sphere and the *diametrically opposite* point. A phosphate should have a *large negative* CC and three approximately equal moments of density (for a sphere of radius 2.5 Å about the phosphorus atom).
The sphere of influence algorithm

The variance $V$ of the density on a spherical surface of radius 2.42 Å is calculated around each pixel in the map. This spherical surface was chosen because 2.42 Å is a typical 1,3-distance in proteins and DNA. A high $V$ indicates that a pixel could correspond to a true atomic position.

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

For pixels 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 for positive density in the CCP4 program ACORN and (like ACORN) works best with high resolution data ($d < 2$ Å).

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

An empirical weighting scheme for phase recombination is used to combat model bias.
Although SHELXE makes no use of histogram matching, the **sphere of influence algorithm** is able to bring the histogram much closer to the one for the correct structure!
The free lunch algorithm

The *free lunch algorithm* (FLA) is an attempt to extend the resolution of the data by including, in the density modification, reflections at higher resolution than have been measured. It was first published by the Bari group in 2005 but appears to have been implemented earlier by the York group in ACORN.

The unexpected conclusion was that if these phases are now used to recalculate the density, using very rough guesses for the (unmeasured) amplitudes, the density actually improves! The FLA is incorporated in SHELXE and tests confirm that the phases of the observed reflections improve, at least when the native data have been measured to a resolution of 2 Å or better.
Maps before and after a free lunch

Best experimental phases after density modification (MapCC 0.57)

After expansion to 1.0 Å with virtual data (MapCC 0.94)

Why do we get a free lunch?

It is not immediately obvious why inventing extra data improves the maps. Possible explanations are:

1. The algorithm corrects Fourier truncation errors that may have had a more serious effect on the maps than we had realised.

2. Phases are more important than amplitudes (see Kevin Cowtan’s ducks and cats!), so as long as the extrapolated phases are OK any amplitudes will do.

3. Zero is a very poor estimate of the amplitude of a reflection that was not measured.
The SHELXE autotracing algorithm

A fast but very crude autotracing algorithm has been incorporated into the density modification in SHELXE. *It is primarily designed for iterative phase improvement starting from very poor phases.* The tracing proceeds as follows:

1. Find potential \( \alpha \)-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 of symmetry elements or existing 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 ($\phi/\varphi$ 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):

\[
\begin{align*}
\text{C} & \quad \text{N} - \quad \text{H} & \quad \text{O} \\
\text{C} & \quad \text{N} - \quad \text{H} & \quad \text{O} \\
\text{C} & \quad \text{N} - \quad \text{H} & \quad \text{O} \\
\text{C} & \quad \text{N} - \quad \text{H} & \quad \text{O} \\
\end{align*}
\]
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%.
2.0 Å synchrotron data against CuK$_\alpha$ in-house

Test structures (phased on native sulfur plus miscellaneous ions):

A: Apoferritin, B: Insulin, C: Elastase, D: Thaumatin, E: Thermolysin, F: Trigonal trypsin

The data redundancy was rather similar for both sets, varying from about 12 for orthorhombic to 40 or more for cubic. Not included is glucose isomerase, which failed for both approaches.
A surprisingly difficult S-SAD example

We recently investigated the plant thionine Hellethionin D, kindly provided by Franz Kerek who had already determined a respectable-looking NMR structure (PDB 1NBL). With four disulfides in only 46 residues it looked like a sitting duck for long-wavelength sulfur-SAD!

We (Andrea Thorn et al.) collected data from two crystals at a wavelength of 1.9 Å on X12 at EMBL/DESY. From crystal 1 we also collected a short (0.9 Å) wavelength dataset to somewhat higher resolution (1.9 Å) to be used for density modification and structure refinement.
Anomalous statistics for Hellethionin D

Crystal 1 showed most anomalous signal, and is well correlated with crystal 2 (also $\lambda = 1.9$ Å) and the CuKα data from crystal 3. However the CC values with the calculated $\Delta F$ for the final refined structure are rather low.

Even worse, all attempts to locate the sulfurs with SHELXD failed!
Attempted structure solution with MR

Our initial attempts to solve the structure of Hellethionin by molecular replacement using Phaser and the NMR model of the same molecule were equally unsuccessful. Possible explanations are:

1. NMR – Not for Molecular Replacement?
2. Too many molecules in the unit-cell (actually 112)?
3. Our incompetence?

It was possibly a combination of all three; the NMR model had a r.m.s. $C_\alpha$ deviation of 1.34 Å to the final crystal structure. In fact we later discovered that Phaser could solve it easily using a crystal structure model of viscotoxin A1 that has a sequence identity of 51% (r.m.s. $C_\alpha$ deviation 0.76 Å); ironically a structure that we had determined ourselves by sulfur-SAD.
ARCIMBOLDO

ARCIMBOLDO (Isabel Usón) solves structures using PHASER followed by SHELXE with $\alpha$-helices as search fragments. The keys to the success of ARCIMBOLDO are (a) the ability of PHASER to generate many MR solutions (a few of them correct) for very small fragments, (b) the ability of SHELXE to bootstrap from a them to a full trace and provide an unambiguous indication as to which solutions are correct, and (c) massive computer power in the form of a CONDOR cluster.

The successful solutions are recognized by the mean chain length (>10) and the CC for calculated structure factors for the trace against the native data (>25%). Currently native data to 2 Å resolution or better are needed, but otherwise ARCIMBOLDO is very widely applicable because no other model or phase information is required, and has already notched up several spectacular successes.
The solution

The best hit from ARCIMBOLDO had a CC of 25% for the backbone trace and a mean chain length of 16. These figures indicate a potential solution. Rerunning SHELXE for more cycles or running only 5 cycles but correcting the solvent content from 0.45 to 0.55 (7 rather than 8 molecules) improves the CC to 36% and the mean chain length to 34 with about 90% of the residues traced correctly.

Although this model was more than good enough for immediate refinement, we used SHELXE to find 49 ‘sulfur’ atoms via the $\phi-\alpha$ map and used them successfully for SAD phasing. This MRSAD approach has the advantage of eliminating any residual MR model bias.

This solution was refined further with REFMAC. 35 of the ‘waters’ were assigned to $\text{Cl}^-$ and 28 to $\text{Na}^+$ (in some cases with halved occupancies) based on the normal and anomalous maps. The crystals had been grown from ca. 2M NaCl plus smaller concentrations of other salts.
As usual, the internal consistencies of the NMR ensemble (20 molecules shown in brown) and the crystal ‘ensemble’ (7 shown in blue) is much better than the agreement between NMR and crystal. The CYS residues are numbered.
Conclusions

Clearly, the 2 molar Cl\(^-\) (and possibly other) ions in the disordered solvent are making the heavy atom location much more difficult. So for sulfur-SAD phasing one should NEVER crystallize from NaCl or (NH\(_4\))\(_2\)SO\(_4\) solution! The poor CC values between the \(\Delta F\) values at low resolution may also be indicative of further disordered anomalously scattering solvent.

If the crystals refuse to grow under any other conditions, searching for many sites for a long time might lead to a solution. Alternatively MRSAD is less sensitive to the number of anomalous scatterers, so even a poor MR solution could be used to find the sulfur and most of the chloride sites.

And there is always ARCIMBOLDO!
Acknowledgements

I am particularly grateful to Andrea Thorn and Isabel Usón who did most of the work on Hellethionin, Franz Kerek for the sample of Hellethionin, Tim Grüne, Christian Grosse and Manfred Weiss for help with data collection on X12, Isabel Usón and Dayté Rodríguez for running Arcimboldo, and Tobias Beck for the I3C data.

