The future of direct methods

IUCr Computing School, Siena,
August 2005

George M. Sheldrick

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

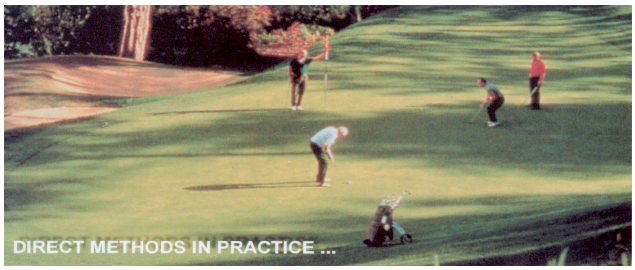
The crystallographic phase problem

- In order to calculate an electron density map, we require both the intensities $I = |F|^2$ and the phases ϕ of the reflections *hkl*.
- The information content of the phases is appreciably greater than that of the intensities.
- Unfortunately, it is almost impossible to measure the phases experimentally !

This is known as the *crystallographic phase problem* and would appear to be difficult to solve!

Despite this, for the vast majority of small-molecule structures the phase problem is solved routinely in a few seconds by black box *direct methods*.

Finding the minimum



The Sayre equation

Sayre (1952). In the same issue of Acta Cryst., Cochran and Zachariasen independently derived phase relations and showed that they were consistent with Sayre's equation:

$$F_h = q \sum_{h'} (F_{h'} F_{h-h'})$$

where q is a constant dependent on $\sin(\theta)/\lambda$ for the reflection h (hkl) and the summation is over all reflections h' ($h'k'l'$) . Sayre derived this equation by assuming equal point atoms. For such a structure the electron density (0 or Z) is proportional to its square (0 or Z^2) and the *convolution theorem* gives the above equation directly.

The Sayre equation is (subject to the above assumptions) exact, but requires complete data including F_{000}.

Normalized structure factors

Direct methods turn out to be more effective if we modify the observed structure factors to take out the effects of atomic thermal motion and the electron density distribution in an atom. The normalized structure factors E_h correspond to structure factors calculated for a point atom structure.

$$E_h^2 = (F_h^2/\varepsilon) / \langle F^2/\varepsilon \rangle_{\text{resl. shell}}$$

where ε is a statistical factor, usually unity except for special reflections (e.g. 00*l* in a tetragonal space group). $\langle F^2/\varepsilon \rangle$ may be used directly or may be fitted to an exponential function (Wilson plot).

The tangent formula (Karle & Hauptman, 1956)

The tangent formula, usually in heavily disguised form, is still a key formula in small-molecule direct methods:

$$\tan(\phi_h) = \frac{\sum_{h'} | E_{h'} E_{h-h'} | \sin(\phi_{h'} + \phi_{h-h'})}{\sum_{h'} | E_{h'} E_{h-h'} | \cos(\phi_{h'} + \phi_{h-h'})}$$

The sign of the sine summation gives the sign of $\sin(\phi_h)$ and the sign of the cosine summation gives the sign of $\cos(\phi_h)$, so the resulting phase angle is in the range 0-360°.

The Multan Era (1969-1986)

The program MULTAN (Woolfson, Main & Germain) used the tangent formula to extend and refine phases starting from a small number of reflections; phases were permuted to give a large number of starting sets. This multisolution (really multiple attempt) direct methods program was user friendly and relatively general, and for the first time made it possible for non-experts to solve structures with direct methods. It rapidly became the standard method of solving small-molecule structures.

Yao Jia-Xing (1981) found that it was even better to start from a large starting set with random phases (RANTAN), and this approach was adopted by most subsequent programs.

Negative quartets: using the weak data too

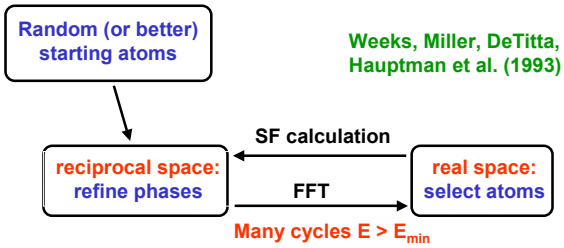
Schenk (1973) discovered that the quartet phase sum:

$$\Phi = \phi_h + \phi_{h'} + \phi_{h''} + \phi_{-h-h'-h''}$$

is, in contrast to the TPR sum, more often close to 180° than 0° when the four primary E-values E_h, $E_{h'}$, $E_{h''}$ and $E_{-h-h'-h''}$ are relatively large and the three independent cross-terms $E_{h+h'}$, $E_{h+h''}$ and $E_{h'+h''}$ are all small. Hauptman (1975) and Giacovazzo (1976) derived probability formulas for these negative quartets using different approaches; Giacovazzo's formula is simpler and more accurate and so has come into general use.

Although this phase information is weak (and depends on $1/N$ rather than $1/N^{½}$ for TPRs) tests based on negative quartets discriminate well against uranium atom false solutions.

Dual space recycling



Weeks, Miller, DeTitta, Hauptman et al. (1993)

If the figures of merit indicate a solution, it can be expanded to the complete structure using all data

Implemented in SnB and (later) SHELXD

The correlation coefficient between E_o and E_c

$$CC = \frac{100\,[\,\Sigma(wE_o\,E_c)\,\Sigma w - \Sigma(wE_o)\,\Sigma(wE_c)\,]}{\{\,[\Sigma(wE_o^{\,2})\Sigma w - (\Sigma wE_o)^2]\cdot[\Sigma(wE_c^{\,2})\Sigma w - (\Sigma wE_c)^2]\,\}^{½}}$$

Fujinaga & Read, *J. Appl. Cryst.* 20 (1987) 517-521.

For data to *atomic resolution*, a CC of 65% or more almost always indicates a correct solution.

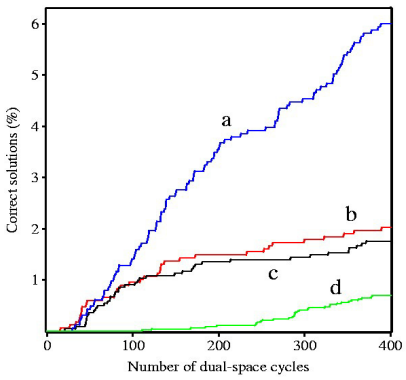
Strategies for atom selection

- Simply keep top N atoms
- Eliminate atoms to maximize e.g. $\Sigma\, E_c^2(E_o^2\text{-}1)$
- Eliminate 30% atoms at random

Strategies for phase refinement

- Do no phase refinement
- Reduce the minimal function by the parameter-shift method
- Fix 30-50% of the phases with largest E_c, derive the rest by tangent expansion

Gramicidin A (N=317) - different strategies



a: random omit + tangent expansion

b: random omit + minimal function

c: top N peaks + minimal function

d: random omit + no phase refinement

Random OMIT maps

Omit maps **are 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 is the use of *random omit maps* in the search stage. About 30% of the peaks are omitted at random and the phases calculated from the rest are refined. The resulting phases and observed E-values are used to calculate the next map, followed by a peaksearch. This procedure is repeated 20 to 500 times.

Although the random omit and probabilistic Patterson sampling appreciably improve the efficiency of direct methods, using both together is not much better than either alone. Usually we use the probabilistic Patterson sampling for the location of heavy atoms for macromolecular phasing and random omit maps for *ab initio* structure solution.

Unknown structures solved by SHELXD

Compound	Sp. Grp.	N(mol)	N(+solv)	HA	d(Å)
Hirustasin	$P4_32_12$	402	467	10S	1.20
Cyclodextrin	$P2_1$	448	467		0.88
Decaplanin	$P2_1$	448	635	4Cl	1.00
Cyclodextrin	P1	483	562		1.00
Bucandin	C2	516	634	10S	1.05
Amylose-CA26	P1	624	771		1.10
Viscotoxin B2	$P2_12_12_1$	722	818	12S	1.05
Mersacidin	$P3_2$*	750	826	24S	1.04
Feglimycin	$P6_5$*	828	1026		1.10
Tsuchimycin	P1	1069	1283	24Ca	1.00
rc-WT Cv HiPIP	$P2_12_12_1$	1264	1599	8Fe	1.20
Cytochrome c3	$P3_1$	2024	2208	8Fe	1.20

*twinned

The 1.2 Å rule

"Experience with a large number of structures has led us to formulate the empirical rule that if fewer than half the number of theoretically measurable reflections in the range 1.1-1.2 Å are "observed", it is very unlikely that the structure can be solved by direct methods" **[Sheldrick, 1990].**

Morris & Bricogne, Acta Cryst. D59 (2003) 615-617 gave an explanation: the variation of the experimental E^2 with resolution shows that data in the range 1.2-1.0 Å have a higher *information content*.

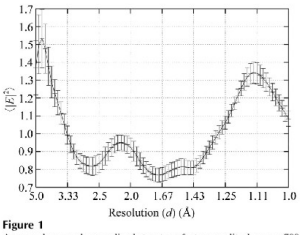


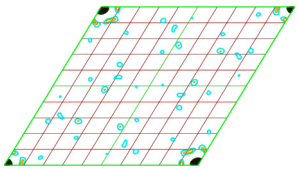
Figure 1
Averaged squared normalized structure-factor amplitudes over 700 protein structures with standard deviations calculated from the population of individual $|E|^2$ profiles.

Heavy atoms and the 1.2 Å rule

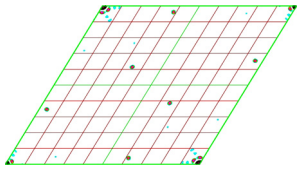
When heavier atoms such as S or Fe are present, this rule can be relaxed a little. Tests using high resolution data artificially truncated (or not measured) to a resolution worse than the diffraction limit of the crystal also tend to perform better. Many of the largest structures solved by direct methods fall into these categories.

When heavy atoms are present, probabilistic sampling of a *super-sharp Patterson* [e.g. with coefficients $\sqrt{(E^3F)}$] is a good way to kick-start *ab initio* direct methods.

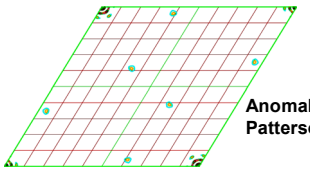
Cytochrome c6 Patterson



Normal native Patterson



Super-sharp native Patterson



Anomalous difference Patterson

Resolving the resolution problem

Replacing peak picking by some form of density modification, as used by Giacovazzo et al. in SIR2003 and in ACORN (Yao Jia-Xing et al.) appears to alleviate the resolution problem a little, though maybe only to 1.3 or 1.4 Å. Recent versions of SIR (Giacovazzo et al.) make extensive use of iterative density modification, e.g. using only the strongest *E*-values and setting all but the highest density to zero. To judge from the published tests, SIR2005 may be more effective than SHELXD and SnB for large structures at borderline atomic resolution, especially when heavier atoms are present.

A more far-reaching solution will probably be to find a clever way of exploiting chemical information that is not too expensive in terms of computer time. It should be noted that searching for small fragments instead of single atoms is particularly slow. We have successfully taken a small step in this direction by fitting S_2-units when locating the anomalous scatterers for sulfur-SAD phasing.

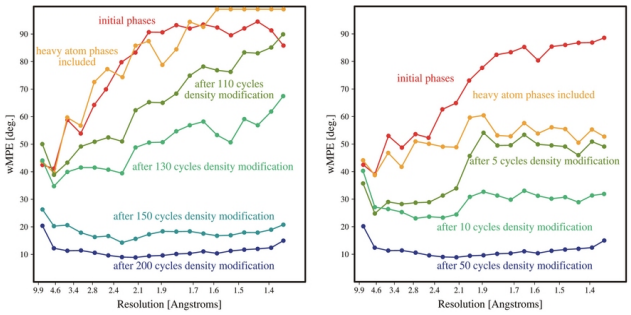
Disulfide bond resolution

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 small fragments in the real space part of the dual-space recycling may be a good way of extending direct methods to lower resolution, provided that it can be done efficiently.

Mean phase errors as a function of resolution for SAD phasing of cubic insulin at a wavelength of 1Å



without disulfide resolution | with disulfide resolution

Other promising methods for atomic resolution data

1. Iterated projections [Elser, *Acta Cryst.* A59 (2003) 201-209]. This is a complicated iterative density modification algorithm that seems to be quite effective and not much slower than SnB or SHELXD. So far it is restricted to space group P1.
2. Charge flipping [Oszlanyi & Suto, *Acta Cryst.* A60 (2004) 134-141; A61 (2005) 147-152; Wu et al., A60 (2004) 326-330]. This algorithm is simple and easy to program, but in my tests was not quite as effective as the almost as simple random omit method (on its own without the tangent formula etc.).
3. Integer programming [Viai & Sahinidis, *Acta Cryst.* A59 (2003) 452-458 & A61 (2005) 445-452]. By reducing CENTROSYMMETRIC direct methods to an *integer programming problem*, it appears that these authors have indeed found a solution to the phase problem *in polynomial time*. Tests have shown that this method is as fast or faster, and less likely to fail, than existing methods for centrosymmetric structures.

Ab initio direct methods at lower resolution

Considerable progress has also been made at very low resolution, where the number of reflections is so small that it is feasible to test many phase permutations [e.g. Lunina, Lunin & Urzhumtsev, *Acta Cryst.* D59 (2003) 1702-1713]. Solutions are selected on the basis of good connectivity and the right number of connected fragments in the cell, and then merged with each other. In favourable cases it may be possible to begin to see secondary structure in the 4 to 6 Å maps that are produced. Such methods are very sensitive to missing or wrongly measured low order reflections.

Conditional optimisation [Scheres & Gros, *Acta Cryst.* D57 (2001) 1820-1828] is a sort of molecular dynamics with N atoms in a box subject to a very general force field so that chemically sensible ensembles of atoms are favoured. In principle this is a promising method, but will probably require massive computer resources.

Structure solution in P1

It has been observed [e.g. Sheldrick & Gould, *Acta Cryst.* B51 (1995) 423-431; Xu et al., *Acta Cryst.* D56 (2000) 238-240; Burla et al., *J. Appl. Cryst.* 33 (2000) 307-311] that it may be more efficient to solve structures in P1 and then search for the symmetry elements later. This works particularly well for solving $P\bar{1}$ structures in P1.

I thought that this might be a good way of tackling problems where the space group is not clear. A decision as to the space group could simply be postponed until the structure has been solved! However after much effort I have come to the conclusion that, although the approach works well in straightforward cases, in pseudosymmetry cases there may be a problem in recognising the correct solution to the phase problem, so the current procedure of trying all possible space groups may be more effective!

More than 90% of the algorithms I have devised and programmed turned out, on objective assessment, not to represent improvements on current practice. This was simply one more example.

Acknowledgements

I am particularly grateful to Isabel Usón, Thomas R. Schneider, Stephan Rühl and Tim Grüne for many discussions.

SHELXD: Usón & Sheldrick (1999), *Curr. Opin. Struct. Biol.* 9, 643-648; Sheldrick, Hauptman, Weeks, Miller & Usón (2001), *International Tables for Crystallography Vol. F*, eds. Arnold & Rossmann, pp. 333-351; Schneider & Sheldrick (2002), *Acta Cryst.* D58, 1772-1779.

SHELXE: Sheldrick (2002), *Z. Kristallogr.* 217, 644-650; Debreczeni, Bunkóczi, Girmann & Sheldrick (2003), *Acta Cryst.* D59, 393-395; Debreczeni, Bunkóczi, Ma, Blaser & Sheldrick (2003), *Acta Cryst.* D59, 688-696; Debreczeni, Girmann, Zeeck, Krätzner & Sheldrick (2003), *Acta Cryst.* D59, 2125-2132.

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