Classical direct methods

Göttingen, December 4th 2008

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 $hk\ell$.**
- **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! (*np-hard*?)

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*.

The Sayre equation

In the same issue of Acta Cryst. (1952), Sayre, Cochran and Zachariasen independently derived *phase relations* and showed that they were consistent with the equation now known as the Sayre 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.

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

where ε is a statistical factor, usually unity except for special reflections (e.g. 00ℓ 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 a 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(ϕ_h) and the sign of the cosine summation gives the sign of cos(ϕ_h), so the resulting phase angle is in the range 0-360º.

The triple phase relation

If we express the Sayre equation in terms of *E* :

$$E_h = q \sum_{h'} (E_{h'} E_{h-h'})$$

and compare the phases of the left and right hand sides of one term in the summation, we obtain:

$$\phi_h = \phi_{h'} + \phi_{h-h'} \quad \text{(modulus 360º)}$$

By means of statistical assumtions, for example that the structure consists of equal resolved point atoms, we can obtain a probability distribution (Cochran, 1955) for the error in this TPR:

$$P(\Phi) = g \exp(2 |E_h E_{h'} E_{h-h'}| / N^{½})$$

where $\Phi = \phi_h - \phi_{h'} - \phi_{h-h'}$, *g* is a normalizing factor, and *N* is the number of (equal) atoms in the (primitive) unit-cell.

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.

Disadvantages of the tangent formula

- **In space groups without translational symmetry (e.g. $P\bar{1}$, C2/m, R3) the tangent formula fits exactly to the trivial but wrong (uranium atom) solution with all phases zero.**
- **Enantiomer information tends to be lost in polar space groups (e.g. P1, C2) resulting in a pseudo-centrosymmetric false solution.**
- **Heavy atoms tend to be emphasized at the expense of the light atoms (Sayre's equation can be regarded as a 'squaring equation').**
- **Only the strongest ca. 15% of the data are used.**

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 180º than 0º 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 came into general use.

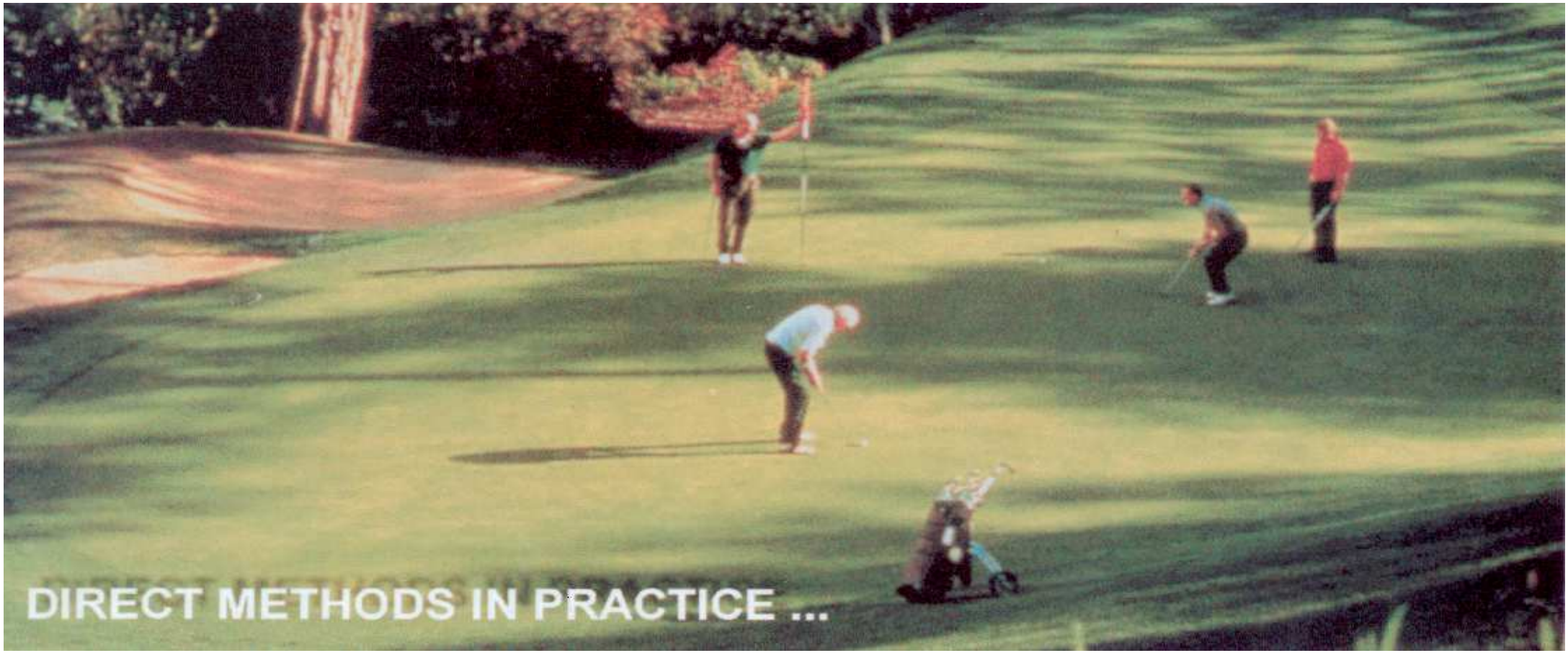
Although this phase information is weak (and depends on $1/N$ rather than $1/N^{½}$ for TPRs) tests based on negative quartets, unlike triplets and the tangent formula, discriminate well against 'uranium-atom' false solutions.

The limits of purely reciprocal space direct methods

Conventional direct methods based on 'improved' versions of the tangent formula and implemented in programs such as MULTAN, RANTAN, SAYTAN, DIRDIF, SIR, SHELXS etc. were extremely efficient at solving small molecule structures up to 100 unique atoms but only succeeded in solving a handful of structures larger than about 200 unique atoms.

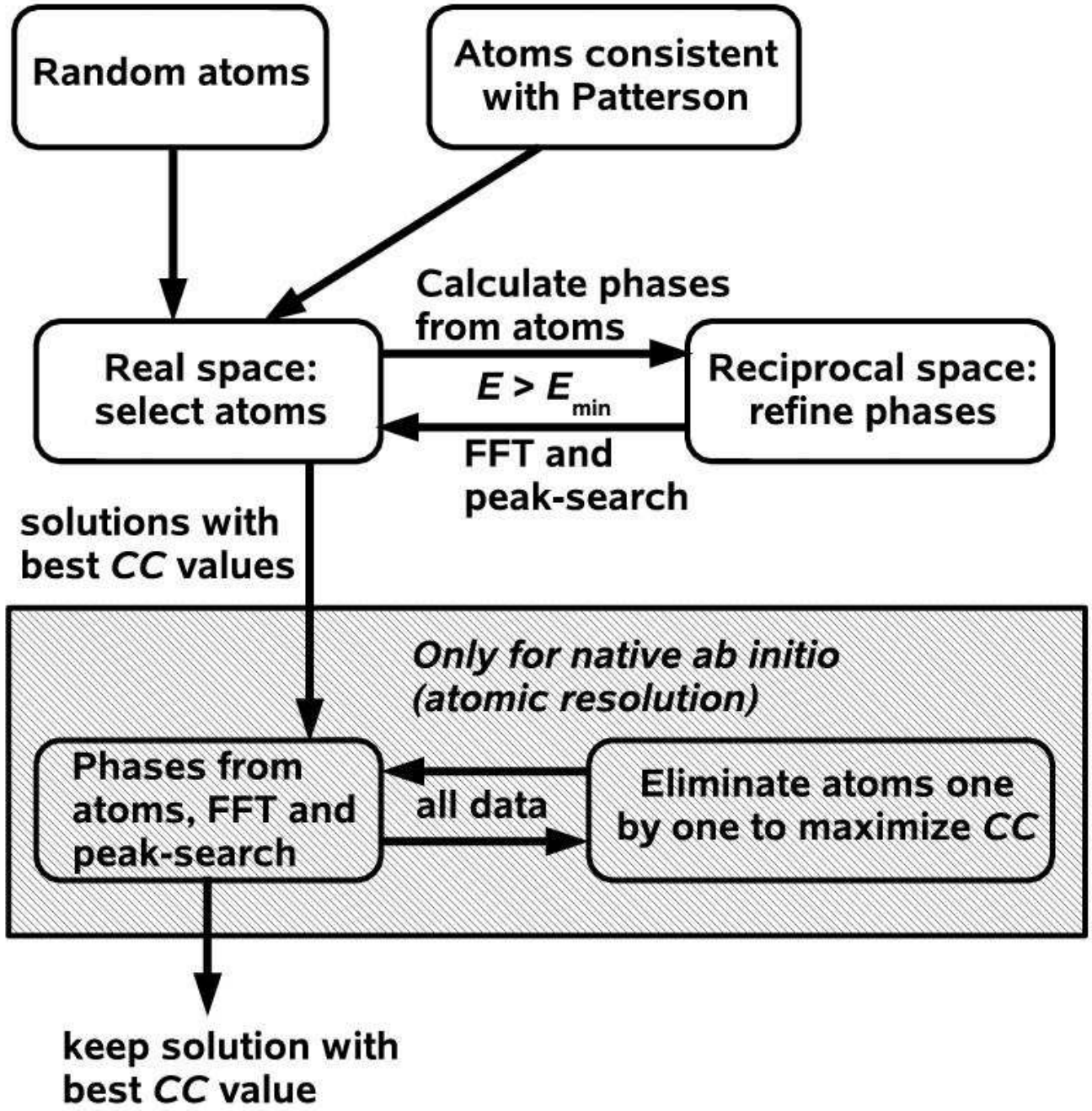
The efficiency of the tangent formula as a phase space search engine lies in its ability to correlate phases widely distributed in reciprocal space. Despite its computational efficiency, the tangent formula tends to lose enantiomorph discrimination and drifts towards *uranium-atom* false solutions, especially for larger structures. It was clearly necessary to *constrain* the phases more tightly to be *chemically reasonable*.

Finding the minimum



Dual space recycling

Introduced with the SnB program by the Buffalo group in 1993 and later used as the basis for SHELXD. The real space part imposes a strong *atomicity* constraint on the phases that are refined in the reciprocal space part. This approach also works well for the location of heavy atoms from SAD, MAD etc. data, because these atoms are also well resolved from each other.



The correlation coefficient between E_o and E_c

$$CC = \frac{100\,[\,\Sigma(wE_o E_c)\,\Sigma w - \Sigma(wE_o)\,\Sigma(wE_c)\,]}{\{\,[\Sigma(wE_o^2)\Sigma w - (\Sigma wE_o)^2] \cdot [\Sigma(wE_c^2)\Sigma w - (\Sigma wE_c)^2]\,\}^{1/2}}$$

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

This is the test used in SHELXD to identify the 'best' solution. For data to *atomic resolution*, a *CC* of 65% or more almost always indicates a correct solution. For heavy atoms from SAD or MAD data, above 30% is probably 'solved'. *CC*(weak), calculated using the reflections NOT used for phasing (like the free *R*) is particularly useful for low resolution SAD or MAD phasing, and should be at least 15%.

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 peak-search. 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.

Fine tuning SHELXD for difficult structures

1. **Use the inner loop (FIND) just to find heavy atoms (with the help of a super-sharp Patterson, PSMF –4) and the outer loop (PLOP) to expand to the full structure.**

2. **If the distribution of peak heights indicates a tendency to produce uranium atom solutions, increase the number of phases held fixed in tangent expansion by increasing the second TANG parameter (e.g. TANG 0.95 0.6).**

3. **Expand the data to P1, then find the true symmetry later (TRIC).**

4. **If the data are borderline for atomic resolution, rename the resulting *name.res* file to *name.ins* and use SHELXE to produce a map (e.g. shelxe name −m20 −s0.4 −e1). This is also a good way of inverting the structure if required (direct methods cannot distinguish enantiomorphs), and also now enables polypeptides to be *autotraced*.**

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 largest substructure solved so far was probably 197 correct Se out of a possible 205 by Qingping Xu of the JCSG (PDB 2PNK).

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]. Remarkably, this rule has withstood the test of time.**

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*.

When heavier atoms are present, this rule can be relaxed somewhat, but then density modification may be needed to complete the structure.

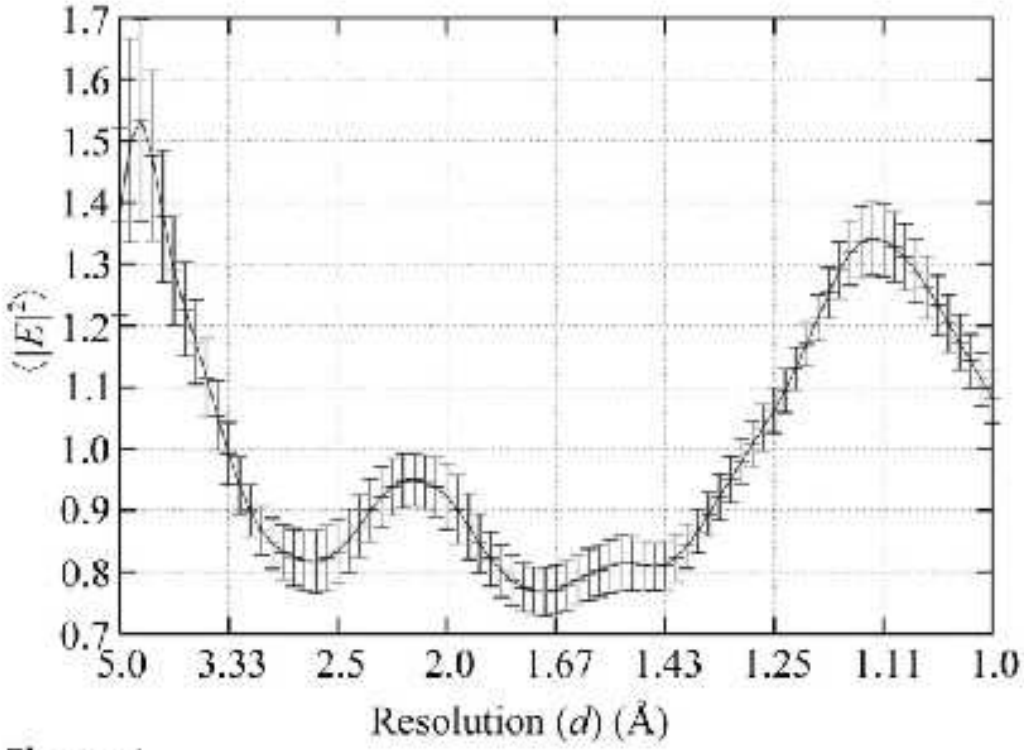


Figure 1
Averaged squared normalized structure-factor amplitudes over 700 protein structures with standard deviations calculated from the population of individual $|E|^2$ profiles.

Strategy for solving a twin with SHELXD

Peter Müller and I have noticed that, even for 50/50 twins without heavy atoms, the SHELXD *inner loop* (FIND) often gives a double image that is strongly biased towards one component.

This has been exploited by incorporating TWIN and (fixed) BASF into the *outer loop* (PLOP). This bistable recycling algorithm 'crystallizes' as one of the two components in the same way that it resolves enantiomorphic double images. More PLOP cycles should be used than for non-twins, and the threshold for entering the outer loop should be lowered.

If heavy atoms are present, the *probabilistic Patterson sampling* PATS should always be used to generate starting atoms, since it is still valid for twins!

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 (most simply by using the TRIC instruction in SHELXD).

This approach requires finding the positions of the symmetry elements (and perhaps with their help determining the space group) *after* solving the structure, for example after solving a $P\bar{1}$ structure in P1 it is necessary to move the complete structure so that an inversion center lies on the origin.

Acknowledgements

I am particularly grateful to Isabel Usón, my group in Göttingen and many SHELX users for discussions and test data.

SHELXD references: 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; Sheldrick (2008), *Acta Cryst.* A64, 112-122.