SHELX Workshop, ACA2007

SHELXD for large small molecules and SHELXC/D for Macromolecular Substructures

Salt Lake City, July 21st 2007

George M. Sheldrick, *Göttingen University*

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

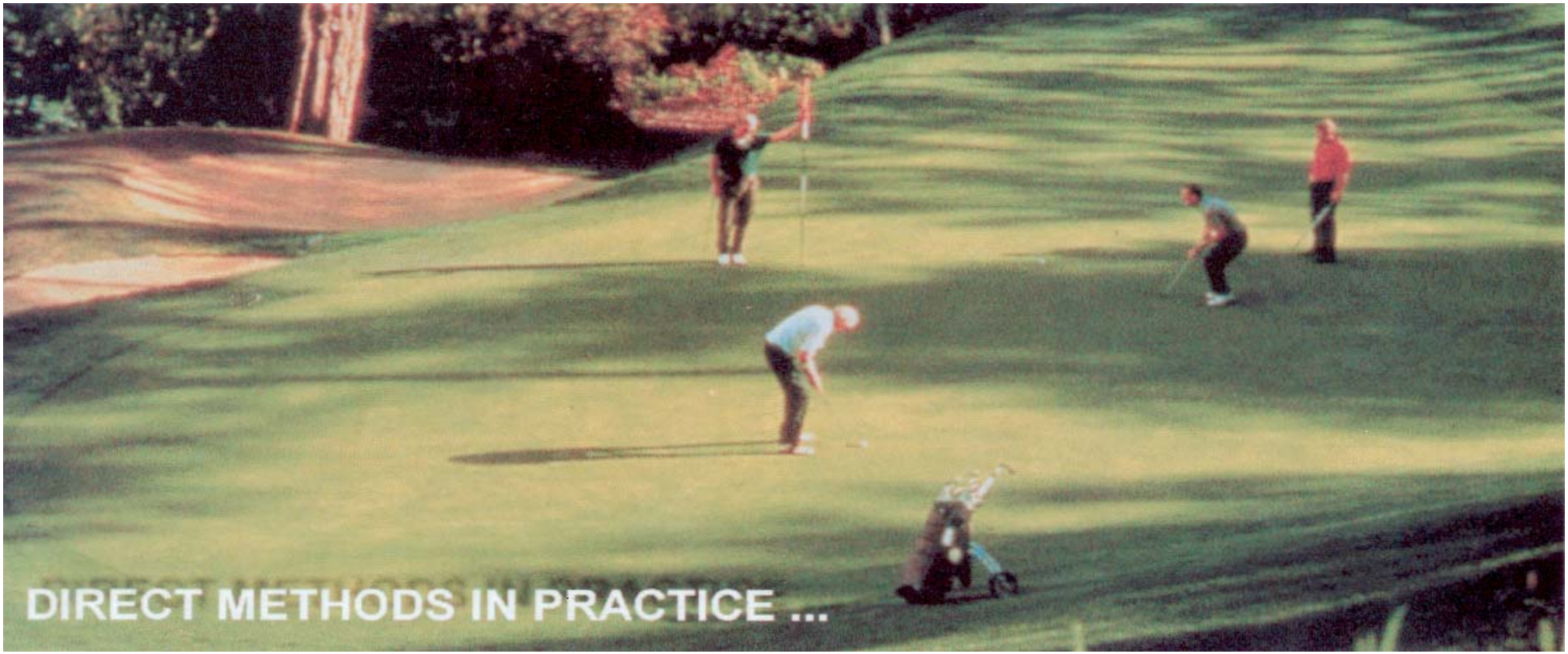
The crystallographic phase problem

- **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



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)

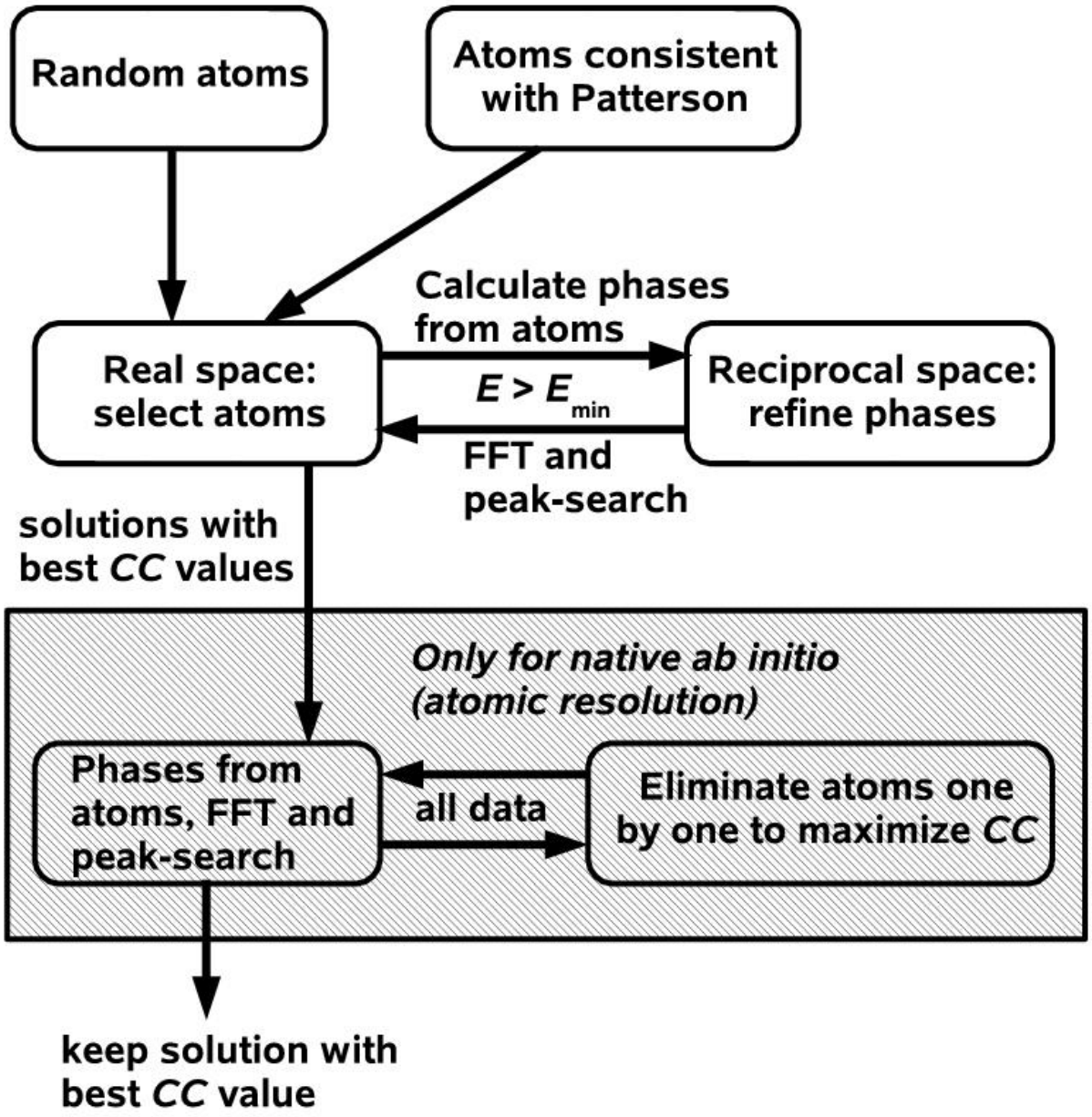
The tangent formula, usually in heavily disguised form, is still a key formula in small-molecule direct methods:

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

The sign of the sine summation gives the sign of sin(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.

Dual space recycling

Introduced with the SnB program by the Buffalo group in 1993. 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.



Probabilistic Patterson Sampling

Each unique general Patterson vector of suitable length is a potential HA-HA vector, and may be employed as a 2-atom search fragment in a translational search based on the *Patterson minimum function*. For each position of the two atoms in the cell, the Patterson height P_j is found for all vectors between them and their symmetry equivalents, and the sum (PSUM) of the lowest (say) 35% of P_j calculated.

It would be easy to find the global maximum of PSUM using a fine 3D grid, but this often does NOT lead to the solution of the structure! A more effective approach is to generate many different starting positions by simply taking the best of a finite number of random trials each time.

The *full-symmetry Patterson superposition minimum function* is used to expand from the two atoms to a much larger number before entering the dual-space recycling.

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.

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%.

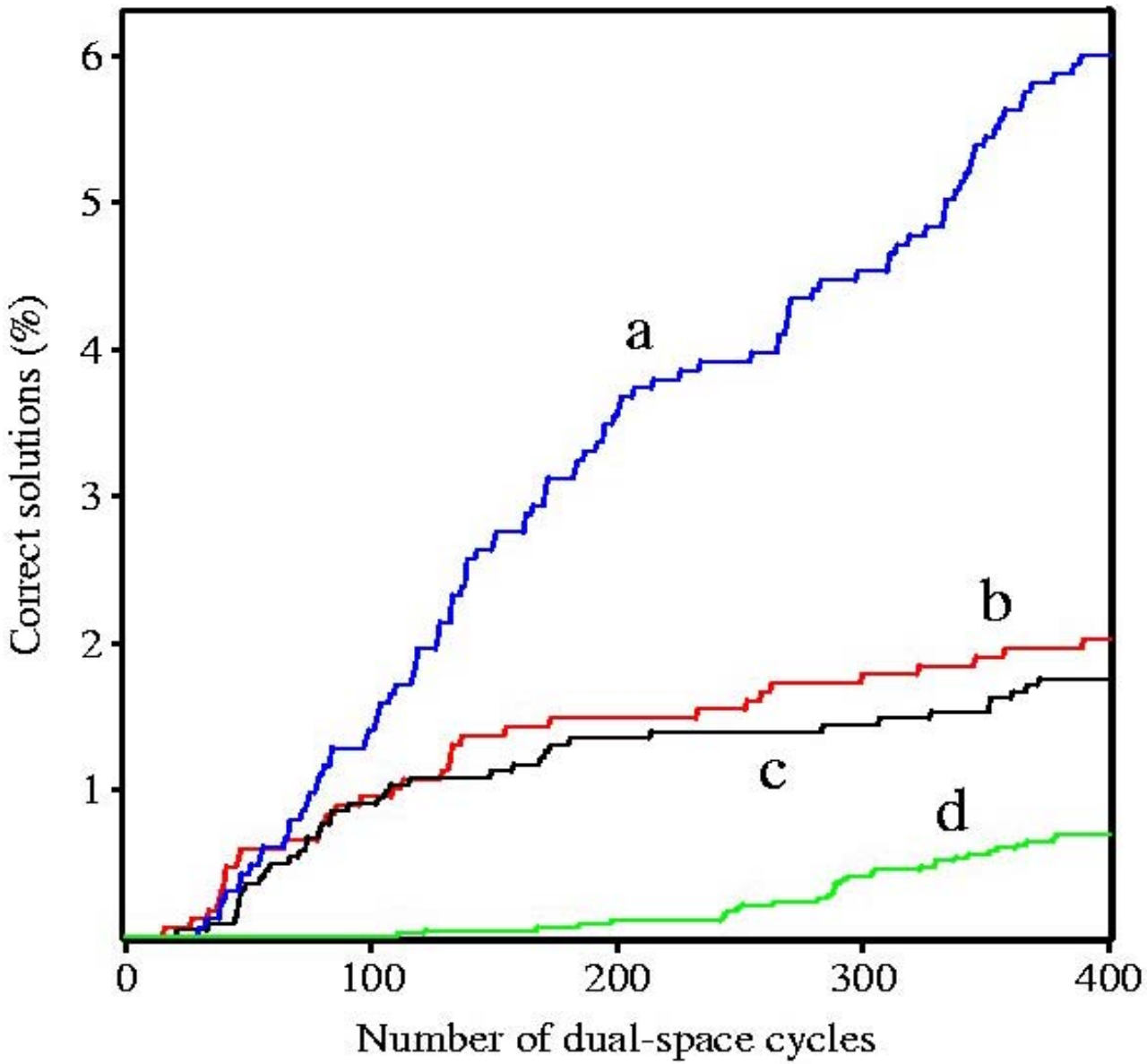
Strategies for atom selection

- Simply keep top *N* atoms
- Eliminate atoms to maximize e.g. $\Sigma E_c^2(E_o^2-1)$
- 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 largest protein substructure solved so far was probably 197 correct Se out of a possible 205 by Qingping Xu of the JCSG (PDB 2PNK).

Experimental phasing of macromolecules

Except in 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 *heavy atoms* to provide *reference phases*. We then calculate the phases ϕ_T of the full structure by:

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

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

1. Location of the heavy atoms.

2. (Refinement of heavy atom parameters and) calculation of ϕ_A.

3. Calculation of starting protein phases using $\phi_T = \phi_A + \alpha$.

4. Improvement of these phases by density modification (and where appropriate NCS averaging).

SAD as a special case of MAD

$$|F_+|^2 = |F_T|^2 + a|F_A|^2 + b|F_T||F_A|\cos\alpha + c|F_T||F_A|\sin\alpha$$

$$|F_-|^2 = |F_T|^2 + a|F_A|^2 + b|F_T||F_A|\cos\alpha - c|F_T||F_A|\sin\alpha$$

where $a = (f''^2 + f'^2)/f_0^2$, $b = 2f'/f_0$, $c = 2f''/f_0$ and $\alpha = \phi_T - \phi_A$

By subtracting the second equation from the first we obtain:

$$|F_+|^2 - |F_-|^2 = 2c|F_T||F_A|\sin\alpha$$

If we assume that the native structure factor $|F_T|$ is given by $|F_T| = \frac{1}{2}(|F_+| + |F_-|)$, this simplifies to:

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

where $|F_A|$ is the heavy atom structure factor) and $\phi_T = \phi_A + \alpha$. Amazingly, this is sufficient to find the heavy atoms and to use them to estimate the protein phases ϕ_T for some reflections.

SAD, SIR, SIRAS and MAD

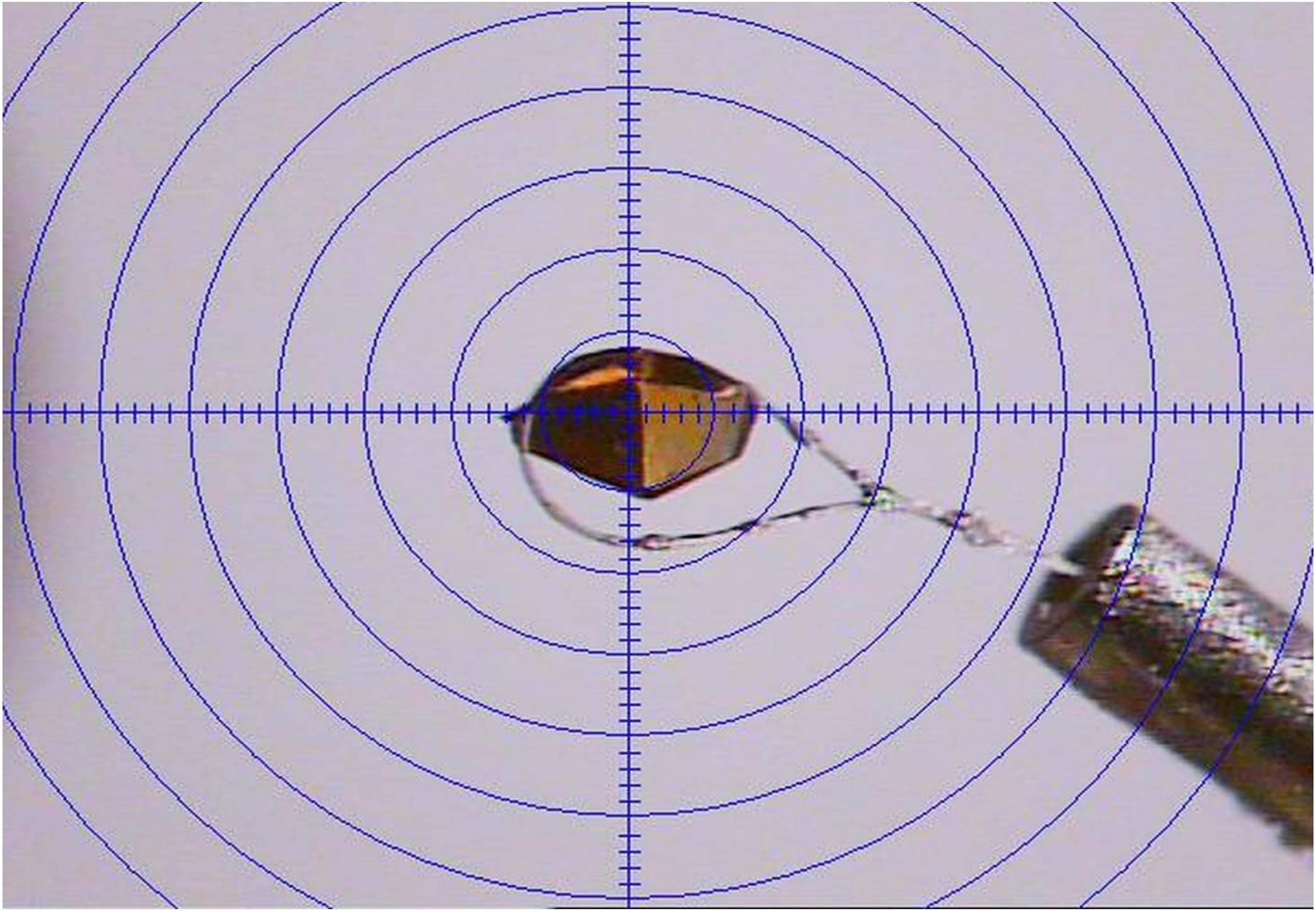
For SAD, the reflections with the largest normalized anomalous differences $|E_A|$ will tend to have α close to 90 or 270º. These reflections are used to find the heavy atoms (only the largest $|E_A|$ are used by direct methods) and to start the phasing.

In the case of SIR, if we assume that the isomorphous difference is small compared to the native structure factor, we obtain the approximation:

$$|F_{deriv}| - |F_{nat}| = b|F_A|\cos\alpha$$

So reflections with large normalized isomophous differences will tend to have α close to 0 or 180º. Although $||F_{deriv}|-|F_{nat}||$ will in general be larger than $||F_+|-|F_-||$, as we shall see α values of 0 or 180º are less useful, and there are problems with lack of isomorphism and scaling.

For MAD (and SIRAS) we have $F_A\sin\alpha$ and $F_A\cos\alpha$ and so we can derive both $|F_A|$ and α.



The importance of the redundancy

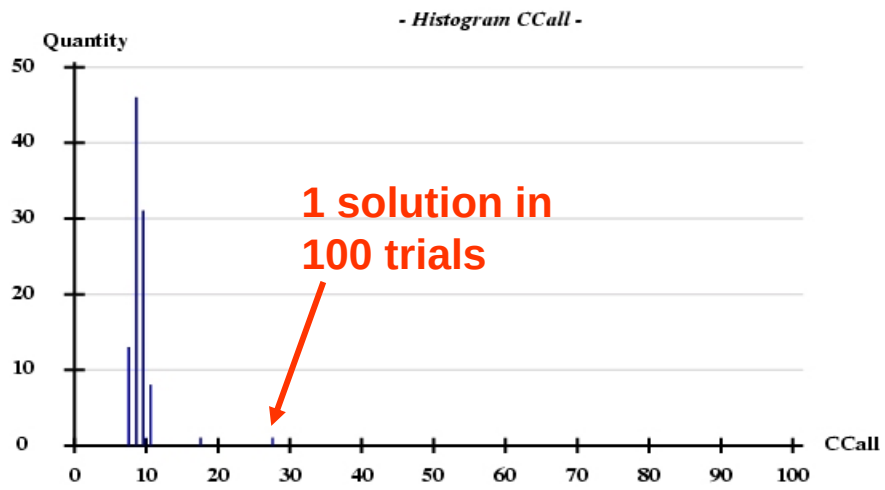
For the cubic insulin data, the redundancy was artificially varied by leaving out scans. As Dauter, Rose, Weiss and many others have demonstrated, the *redundancy* had a major influence on both the success rate and on the discriminating power of the figures of merit.

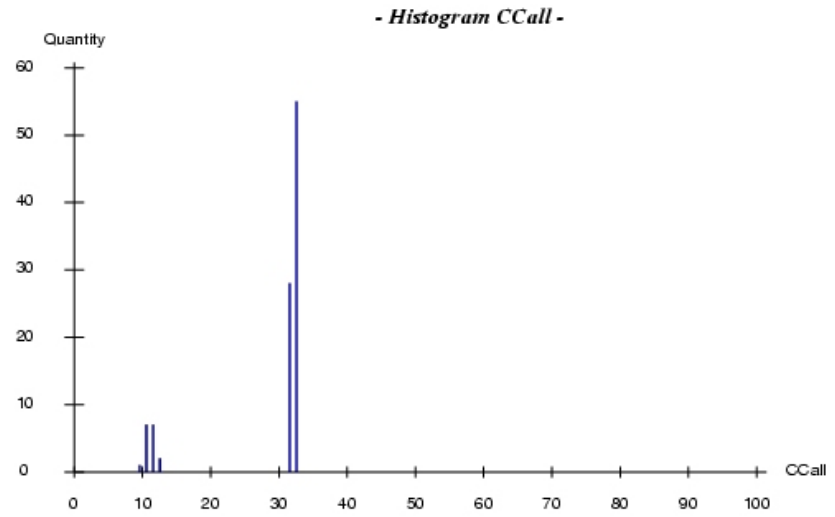
Redundancy (a):	3.8	7.1	14.2	21.4	44.3
Redundancy (b):	2.0	3.7	7.4	11.1	23.0
Hits in 1000 tries:	22	40	222	279	448
CC [%]:	22.7	20.2	28.6	33.9	40.8

(a) Friedels merged, (b) Friedels not merged.

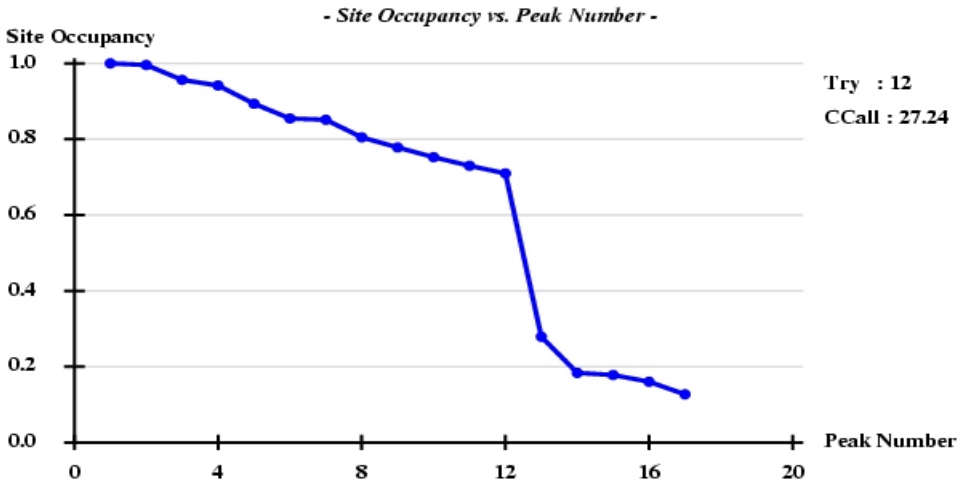
However most such tests were made in-house or on beam-lines of moderate intensity. On third generation synchrotrons *radiation damage* limits the improvement that can be obtained in this way.

SHELXD histograms and occupancies for Elastase

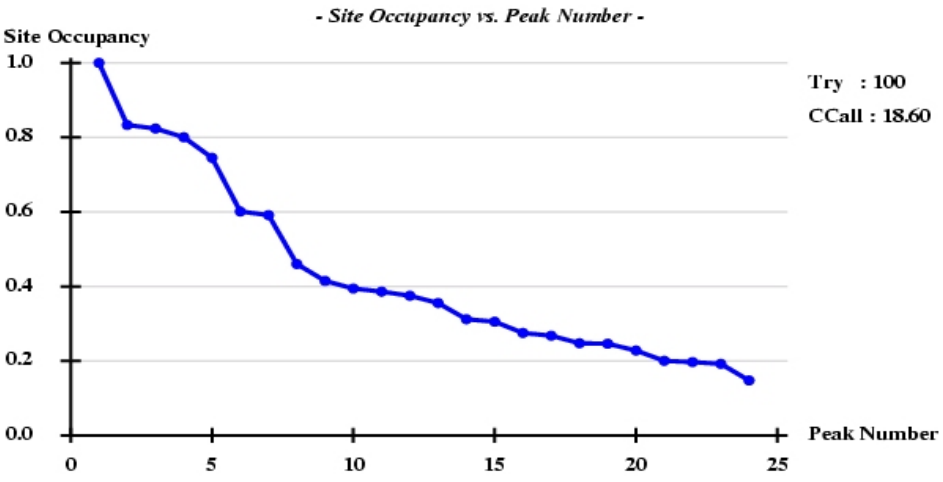




S-SAD



I-SIRAS



The heavy atom enantiomorph problem

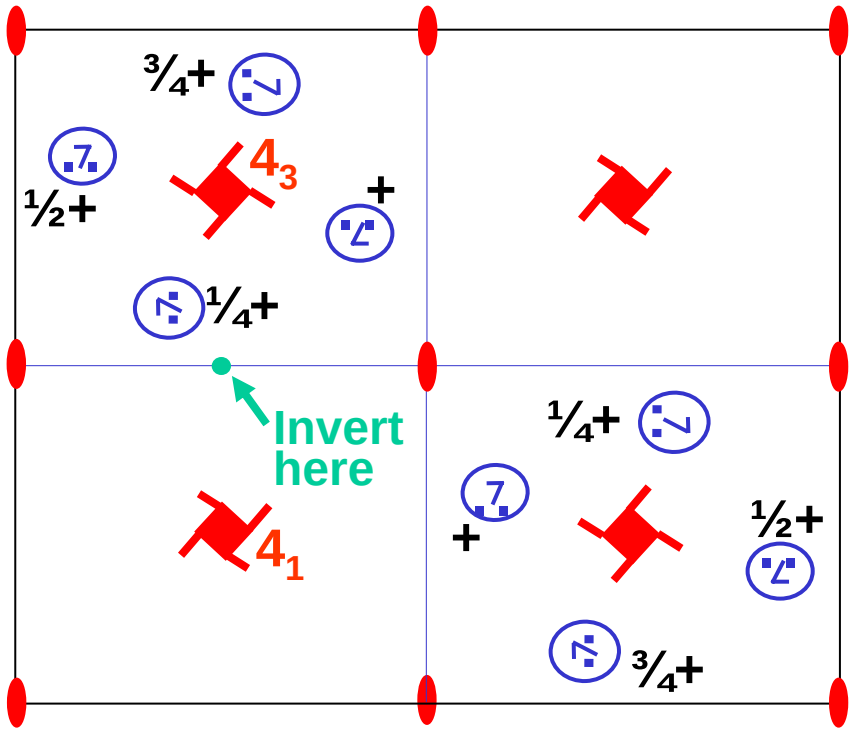
The location of the heavy atoms from the $|F_A|$-values does not define the enantiomorph of the heavy-atom substructure; there is exactly a 50% chance of getting the enantiomorph right. When the protein phases are calculated from the heavy atom reference phases, only one of the two possible maps should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

If the space group is one of an enantiomorphic pair (e.g. $P4_12_12$ and $P4_32_12$) the space group must be inverted as well as the atom coordinates.

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

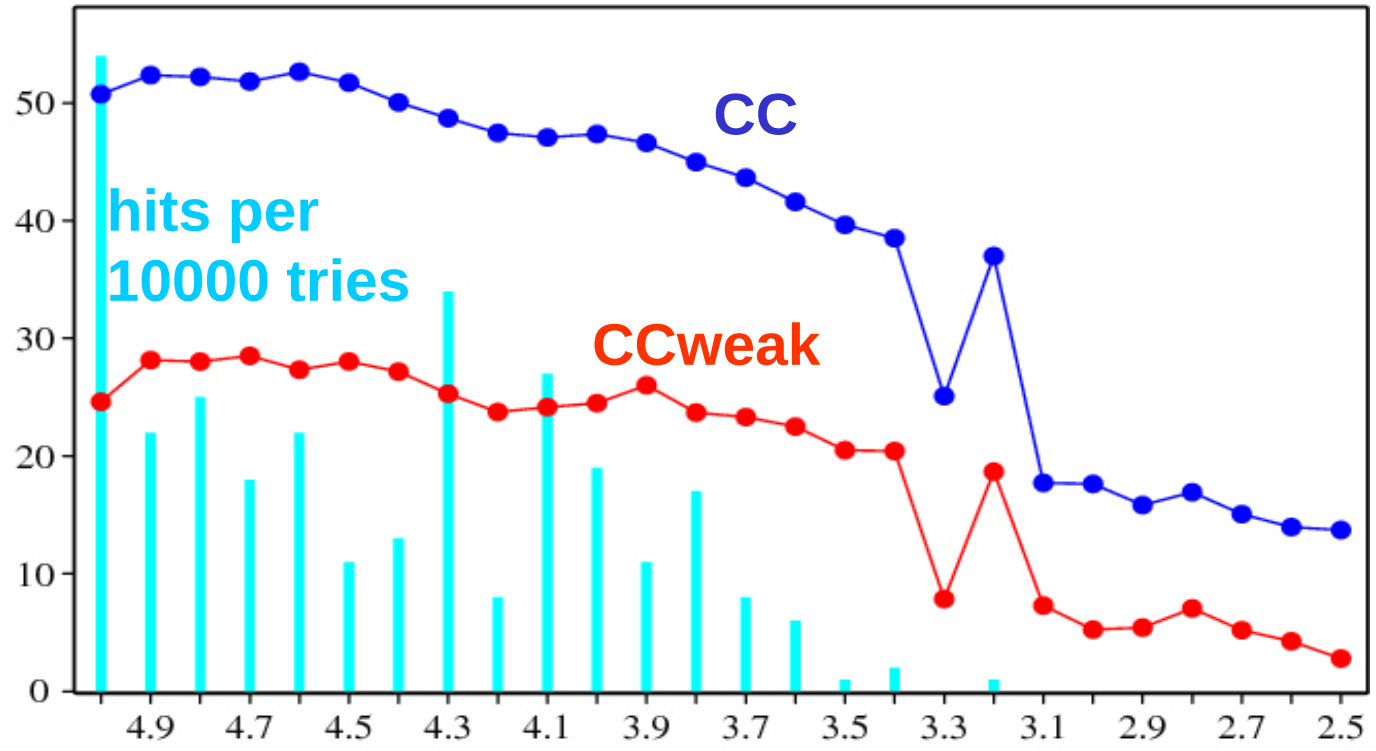
$I4_1$ (1–x,½–y,1–z); $I4_122$ (1–x,½–y,¼–z); $F4_132$ (¼–x,¼–y,¼–z).

Space group $I4_1$



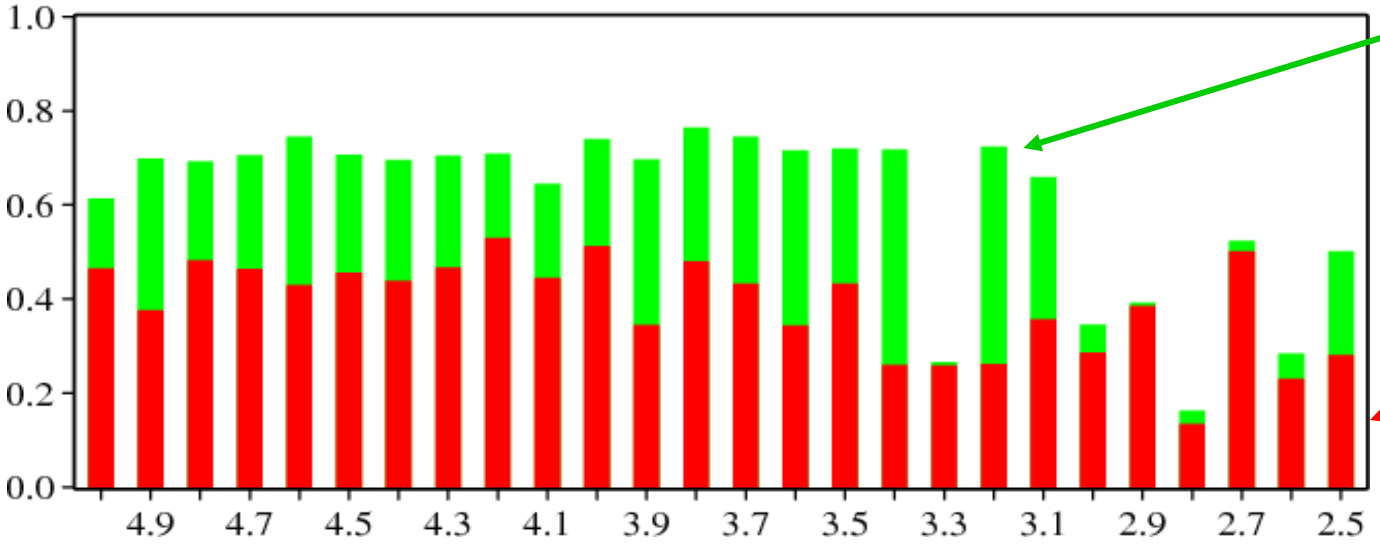
Why do we have to invert a substructure in the space group $I4_1$ in (for example) the point (½,¼,½) rather than in the origin? This problem is a bit like having to change space groups on inverting $P4_1$ to $P4_3$, except that $I4_1$ possesses 4_1 and 4_3 axes and so is its own enantiomorph! Inversion in the origin changes the 4_1 axis to a 4_3 axis and so violates the standard definition of the space group. Inversion in (½,¼,½) not only leaves the symmetry elements as they are, it also inverts the arrangement of the atoms. The helix around the 4_3 axis becomes a helix about a 4_1 axis.

Tendamistat CC, CCweak and hits per 10000 tries



R_{anom} = 0.0270

$R_{p.i.m.}$ = 0.0122



Occupancy (from SHELXD) of peak 8

Occupancy of peak 9

Critical parameters for SHELXD

The Patterson-seeded dual-space recycling in SHELXD is very effective and robust at finding the heavy atom substructure, however attention needs to be paid to:

1. **The *resolution* at which the ΔF-data are truncated, e.g. where the internal CC between the signed anomalous differences of two randomly chosen reflection subsets falls below 30%.**
2. **The *number of sites* requested should be within about 20% of the true value so that the occupancy refinement works well (and reveals the true number).**
3. **In the case of a soak, the rejection of sites on *special positions* should be switched off.**

In difficult cases it may be necessary to fine-tune these settings and run more trials (say 10000 rather than 100).

High throughput pipelines with SHELXC/D/E

The program SHELXC is designed to provide some useful statistical output, in particular to decide at which resolution to truncate the data, and to prepare the files for running SHELXD and SHELXE. All three programs can be run from the command line, but are also intended to be called in from GUIs such as hkl2map.

'*shelxc name*' entered at the command line creates three files:

name.hkl **HKLF4 format merged intensities for density modification with SHELXE (and refinement with SHELXL).**

name_fa.ins **SHELX format instruction file to be read into SHELXD.**

name_fa.hkl **HKLF3 format file for both SHELXD and SHELXE. *h*,*k*,ℓ,*I* and σ*I* are followed by the phase shift α, which is only used by SHELXE.**

The *name_fa.res* output file from SHELXD containing the 'best' heavy atom coordinates is also read into SHELXE.

MAD phasing with SHELXC/D/E

```
shelxc jia <<EOF
NAT jia_nat.hkl
HREM jia_hrem.sca
PEAK jia_peak.sca
INFL jia_infl.sca
LREM jia_lrem.sca
CELL 96.00 120.00 166.13 90 90 90
SPAG C2221
FIND 8
NTRY 10
EOF
shelxd jia_fa
shelxe jia jia_fa -s0.6 -m20
shelxe jia jia_fa -s0.6 -m20 -i
```

Alternatively, one can use a GUI such as Thomas Schneider's hkl2map to guide the selection of file names and parameters!

Acknowledgements

I am particularly grateful to Isabel Usón, Thomas R. Schneider 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.

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