MAD phasing

Göttingen, April 8th 2010

George M. Sheldrick

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

Phasing of macromolecules

Almost all routine small-molecule structure are solved, usually in a few seconds, by the omnipotent *ab initio* direct methods. For macromolecules the following approaches are popular:

1. Molecular replacement (MR). A related structure is used as a search fragment. This works well at low resolution but the final structure suffers from *model bias*, i.e. tends to look more like the search fragment than it should.

2. Multiple isomorphous replacement (MIR). In order to prepare several isomorphous heavy atom derivatives, crystals are soaked in many heavy atom reagents. Although it is never possible to interpret the SIR map from one derivative, the more maps that are averaged together the better the averaged map becomes.

3. Multiple-wavelength anomalous diffraction (MAD). This requires the introduction of e.g. Se, but can in theory give perfect phases.

4. Single-wavelength anomalous diffraction (SAD). The initial uninterpretable map is improved by density modification.

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:

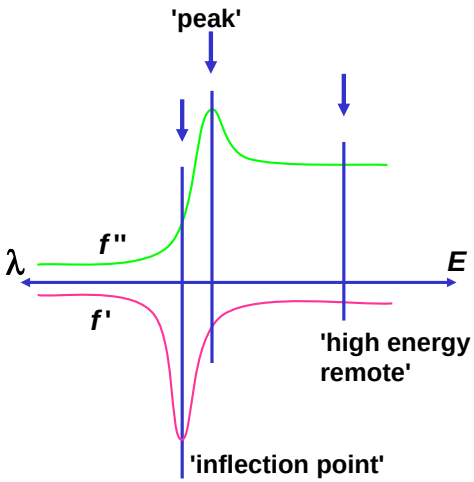
1. Location of the heavy atoms.

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

MAD data



$$f = f_0 + f' + i\,f''$$

MAD data are collected at two or more wavelengths in the vicinity of an absorption edge. A suitable element is selenium with an edge close to 0.98 Å; it can be incorporated into proteins in the form of seleno-methionine. Bromine (bromide soak or bromouracil) and zinc (often present in proteins naturally or in crystallization screens) are also useful.

In a MAD experiment, the small but significant *f"* values and the small differences in *f'* at different wavelengths (up to 6 electrons for Se) are exploited to find phases *experimentally.*

The MAD equations

Karle (1980) and Hendrickson, Smith & Sheriff (1985) showed by algebra that the measured intensities in a MAD experiment should be given by:

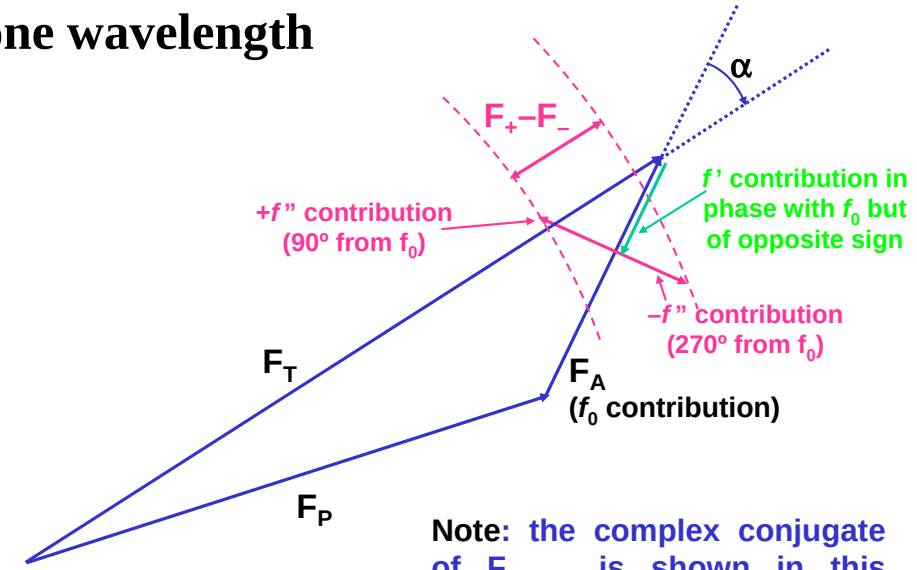
$$|F_+|^2 = |F_T|^2 + a|F_A|^2 + b|F_T||F_A|\cos\alpha + c|F_T||F_A|\sin\alpha$$

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

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

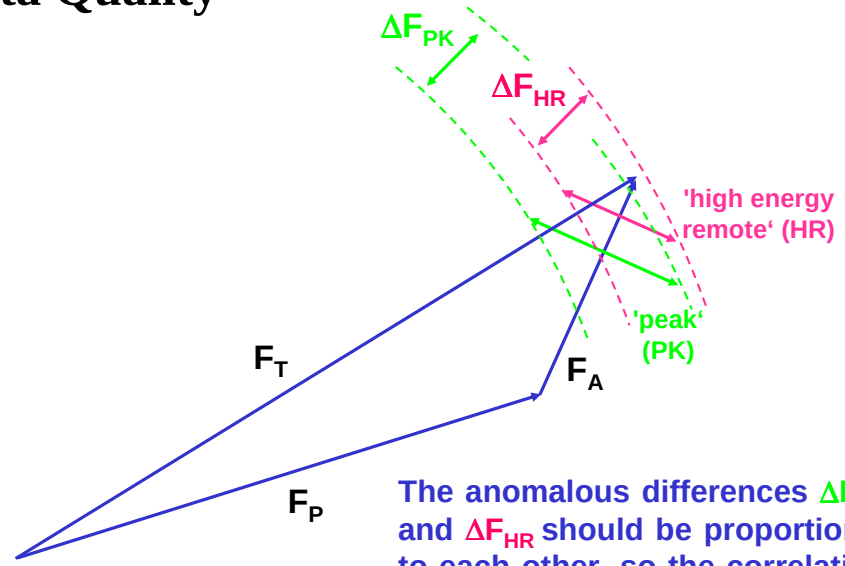
a, *b* and *c* are different for each wavelength. Provided that $|F_\pm|^2$ has been measured at two or more wavelengths we can extract $|F_T|$ (native *F* including heavy atoms but ignoring *f'* and *f"* contributions), $|F_A|$ (heavy atom structure factor) and α (phase shift from heavy atom phase to protein phase) for each reflection. *This works best if the differences in f' and the sum of the f" values are both large.* So for MAD phasing, all we need to do is to use $|F_A|$ to find the heavy atoms, use them to get ϕ_A, and calculate a map with amplitudes $|F_T|$ and phases $\phi_T = \phi_A + \alpha$.

The *f'* and *f"* contributions at one wavelength



Note: the complex conjugate of F_{-h-k-l} is shown in this diagram

Data Quality



The anomalous differences ΔF_{PK} and ΔF_{HR} should be proportional to each other, so the correlation coefficient (CC) between them should be 100% for perfect data.

ΔF Correlation coefficients (%) against resolution (Å)

Apical Domain: 1 x (3 Se-Met in 144) $C222_1$ — Walsh *et al.*, Acta Cryst D55: 1168 (1999)

	Inf	8.0	6.0	5.0	4.0	3.6	3.4	3.2	3.0	2.8	2.6	2.4	2.2
pk	91.2	93.9	93.9	89.6	88.6	89.4	89.4	83.9	76.9	65.7	57.0	44.8	
ip	89.7	90.0	87.0	84.4	79.8	78.9	79.4	74.7	71.1	54.3	47.2	39.2	
lrm	48.5	52.8	52.9	38.0	28.4	34.6	14.2	21.1	24.7	9.1	5.4	-3.7	

RRF: 1 x (4 Se-Met in 185) $P4_32_12$ — Selmer *et al.* Science 286: 2349 (1999)

	Inf	8.0	6.0	5.0	4.6	4.4	4.2	4.0	3.8	3.6	3.4	3.2	3.0
pk	69.3	73.1	62.2	56.9	49.6	45.6	48.6	29.6	20.6	24.6	20.1	14.2	
ip	59.4	58.3	41.9	43.3	40.7	50.4	34.6	24.7	17.5	16.6	8.1	3.9	

Unknown Protein PFH: 4 x (4 Se-Met in 350) $P2_1$ — not to be published

	Inf	8.0	6.0	5.0	4.6	4.4	4.2	4.0	3.8	3.6	3.4	3.2	3.0
pk	33.2	29.5	19.9	10.6	7.7	17.4	7.6	9.8	9.3	13.4	6.0	2.8	
ip	37.6	38.9	37.8	26.5	13.5	24.0	14.2	27.3	25.9	23.1	24.3	22.8	

Thomas R. Schneider

MAD/SAD comparison

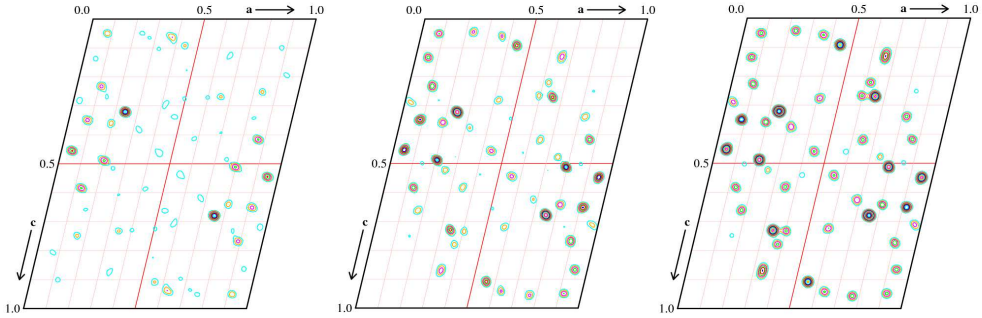
The following tests are based on a high quality 4-wavelength Fe-MAD dataset to 2.9 Å (but the crystal would have diffracted further) provided by Pedro Matias. SHELXD location of the 18 iron atoms gave the following results:

			Occupancies	
	CC	CC_{weak}	Peak 18	Peak 19
SAD, peak only:	58.8	35.7	0.726	0.126
SHELXC 3λ:	71.6	60.3	0.824	0.142
XPREP *f'* and *f"* refined (4λ):	78.4	71.4	0.848	0.082
XPREP MAD on ideal data (3λ):	91.2	87.3	0.867	0.063
F_A calc. directly from atoms:	97.1	96.3	0.966	0.025

In the real world, one would accept CC values of 25/15 for SAD and 40/25 for MAD. The maximum theoretical values are 69 for SAD (Zwart, 2005) and 100 for MAD.

MAD Pattersons

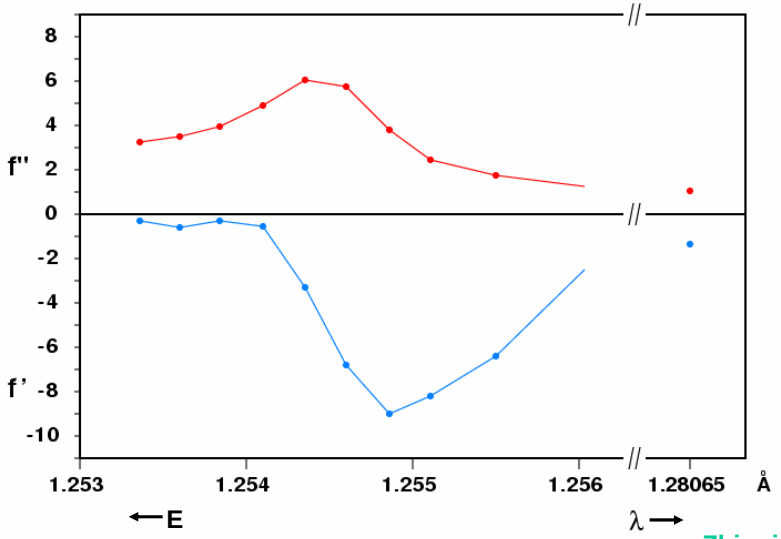
The Harker section at y = ½ for 18 sites in $P2_1$ should show 36 peaks, but several of them overlap and non-Harker peaks may be accidentally close to y = ½:



SAD at peak λ | 4λ MAD (XPREP) | F_A calc. from atoms

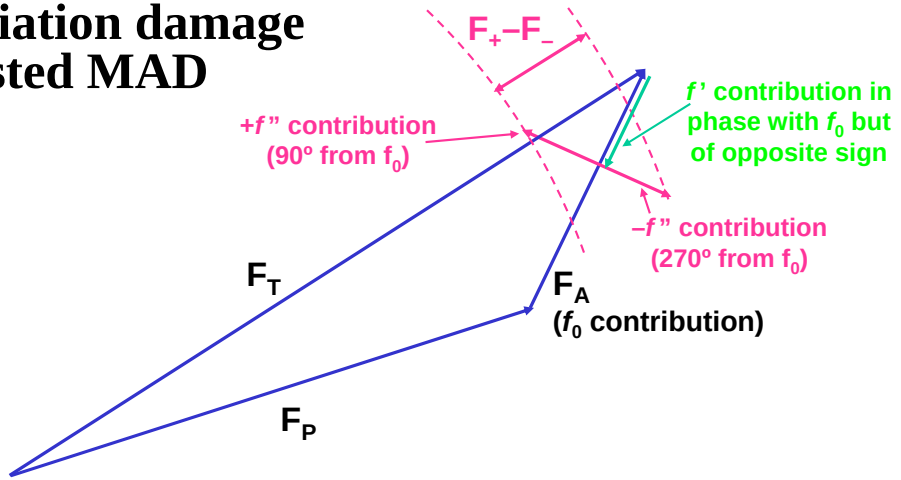
The SAD Patterson uses coefficients $(F_A \sin\alpha)^2$ whereas the other two use F_A^2, so the missing $(F_A \cos\alpha)^2$ appears as noise.

XPREP estimation of *f'* and *f"* data from $[Ta_6Br_{12}]^{2+}$-soaked crystal



Zbigniew Dauter

Radiation damage assisted MAD



If the only effect of radiation damage were to reduce all selenium occupancies for a given wavelength to 0.8, this would cause F_A to be reduced by a factor of 0.8, which would be equivalent to changing *f'* by $-0.2 \times 34 = -6.8$ electrons for that wavelength! This approximately triples the difference in *f'* between the peak and inflection wavelengths, provided that the latter is measured last!

Data collection strategy

In principle, the higher the MoO (multiplicity of observations, previously called redundancy), the more precise the observed intensities and the better the MAD phases should become. Unfortunately, close to the absorption edge the crystal is absorbing the most energy, causing radiation damage.

The best strategy is to collect a good 'peak' dataset with a redundancy of at least 4. If the crystal shows severe radiation damage it may be best to stop there and try SAD phasing. For successful MAD phasing, it is ***ESSENTIAL*** that the ***LAST*** dataset collected with a given crystal is the 'inflection' dataset ***so that the radiation damage and dispersive signals have the same sign***. If radiation damage is minor it may be worth the risk of collecting the 'high energy remote' data in-between, otherwise a fresh crystal can be used for the 'high energy remote' data.

A higher resolution (native) dataset, possibly at lower MoO, is very useful for improving the phases by density modification (and also for the subsequent structure refinement).

When do we need MAD phasing?

Since SAD phasing only requires data at a single wavelength and it is not even necessary to tune this wavelength to be close to the absorption edge, why should we ever do a MAD experiment?

MAD provides about twice as much phase information as SAD, in SAD phasing we rely on the density modification to compensate. This works best for (a) a *high solvent content* (>0.6) or (b) very *high resolution native data* (<1.7 Å).

So, if we have high resolution SeMet or native (e.g. Met instead of SeMet) data, it will be the main factor determining the quality of the map and there is little point in collecting MAD data.

On the other hand, if both the native (if available) and SeMet crystals diffract to worse than say 2.5 Å, we will need all the phase information that we can get, so we should try for MAD. In such cases it is often worth using the MAD phases (without density modification) directly in the (REFMAC) refinement.