Phasing the AP2 core complex with Xe, Hg and Se

Philip Evans

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England

Correspondence e-mail:
pre@mrc-lmb.cam.ac.uk

The 200 kDa core of the heterotetrameric AP2 clathrin adaptor complex was phased with xenon, mercury and selenomethionine derivatives. The phasing has been analysed in retrospect to determine how many of the derivative data sets were necessary for optimum phasing and what features were most useful in deciding which derivatives would make a positive contribution. The relative contributions of the different derivatives indicate that the most important phasing came from the two Xe data sets collected at long wavelengths (1.74 and 1.98 Å) to enhance the anomalous signal. The mercury derivatives were less powerful but made a useful contribution, although inclusion of a poor second wavelength set was detrimental, probably because of radiation damage. The SeMet data were less useful for phasing because of incomplete incorporation of selenium owing to the expression conditions needed, but they were useful in chain-tracing.

1. Introduction

The AP2 clathrin adaptor complex is a heterotetramer involved in the formation of clathrin-coated vesicles. The 200 kDa core of this complex contains the 'trunk' domains of the two large subunits α (residues 1–621 of 938) and $\beta2$ (residues 1–591 of 937), the medium $\mu2$ (435 residues) and small $\sigma2$ (143 residues) subunits. This complex was crystallized in the presence of inositol hexakisphosphate, a mimic of the head group of the lipid phosphatidylinosityl-(4,5)-bisphosphate. The crystals belong to space group $P3_121$, with unit-cell parameters $a = b = 122$, $c = 258$ Å, $\gamma = 120°$ and one complex in the asymmetric unit. The crystals diffract at best to about 2.6 Å resolution, but diffraction is weak beyond about 3.2 Å (Wilson plot B factor about 80 Å^2). The structure was solved at 2.6 Å resolution using a combination of xenon, mercury and SeMet derivatives (Collins *et al.*, 2002). This is a large structure (1790 residues plus inositol hexakisphosphate, with no internal symmetry) and suffered from problems of weak diffraction combined with relatively small phasing signals. These problems were overcome by using high-multiplicity data sets and multiple derivatives to average out the errors. This paper revisits the phasing data to analyse which derivatives were most valuable and whether the strategy adopted was the best.

2. Derivatives and data sets used

During the course of the structure determination, many data sets were collected on both native and derivative crystals. Among the derivatives tried, three were found to be useful: xenon, ethylmercury thiosalicylate (EMTS) and selenomethionine. It was found that including more than one data set of each type of derivative produced apparently better phases,

Table 1
Data-collection statistics.

Values in parentheses are for the highest resolution shell.

Data set	Beamline	Wavelength (Å)	Resolution (Å)	R_{merge}†	R_{meas}‡	Completeness (%)	Multiplicity	$I/\sigma(I)$
Nat6	ID29	0.976	2.9 (3.06)	0.202 (1.005)	0.212 (1.054)	100.0 (100.0)	10.7 (10.9)	11.0 (2.2)
Xe2	ID29	1.743	3.0 (3.16)	0.139 (1.00)	0.150 (1.09)	100.0 (100.0)	13.2 (12.6)	16.2 (2.6)
Xe14	ID29	1.984	3.2 (3.37)	0.132 (1.20)	0.150 (1.34)	100.0 (100.0)	10.1 (9.9)	20.1 (1.7)
EMTS10	EH1	0.934	2.9 (3.06)	0.082 (0.370)	0.110 (0.477)	95.9 (95.9)	3.5 (3.5)	11.9 (2.4)
EMTS7 peak	ID29	1.007	2.9 (3.06)	0.169 (1.74)	0.178 (1.85)	99.9 (99.9)	20.2 (15.6)	17.4 (1.7)
EMTS7 edge	ID29	1.009	2.9 (3.06)	0.134 (1.42)	0.157 (1.75)	99.8 (99.0)	6.7 (5.2)	10.1 (1.0)
Se3 peak	ID29	0.9797	3.1 (3.27)	0.082 (0.358)	0.103 (0.449)	99.2 (100.0)	5.0 (5.1)	16.3 (3.2)
Se3 edge	ID29	0.9798	3.3 (3.47)	0.078 (0.319)	0.104 (0.428)	98.9 (99.8)	3.6 (3.6)	8.7 (2.4)
Se5 peak	ID29	0.9797	3.0 (3.16)	0.090 (0.681)	0.109 (0.845)	99.1 (98.2)	5.7 (4.8)	13.0 (1.7)
Se5 edge	ID29	0.9798	3.0 (3.16)	0.083 (0.624)	0.101 (0.774)	99.1 (98.2)	5.7 (4.8)	14.6 (2.0)
Nat20	EH1	0.834	2.6 (2.74)	0.101 (0.537)	0.109 (0.636)	99.4 (99.4)	5.6 (3.3)	14.8 (2.1)

† $R_{\mathrm{merge}} = \sum\sum_i |I_h - I_{hi}|/\sum\sum_i I_h$, where I_h is the mean intensity for reflection h. ‡ $R_{\mathrm{meas}} = \sum[n/(n-1)]^{1/2}\sum_i|I_h - I_{hi}|/\sum\sum_i I_h$, the multiplicity-weighted R_{merge}.

but not all possible combinations were tried as the phasing calculations in *SHARP* were very slow [the original phase calculations were all performed with *SHARP* version 1.3.12 (de La Fortelle & Bricogne, 1997): the calculations repeated for this paper were performed with the much faster version 2.0.1]. In the end, the phases used to build the model were combined from two separate phasing calculations: one using a native (Nat6), two Xe data sets (Xe2, Xe14) and two EMTS data sets (EMTS10, EMTS7 at two wavelengths, peak and inflection) and the other using two data sets from SeMet crystals (Se3, Se5, each at the peak and inflection points; the remote data sets were discarded owing to excessive radiation damage) together with a native data set (Nat20) treated as a derivative (having Se3 as the 'reference' data set allows better isomorphism between the two SeMet sets). All data were collected at ESRF, mostly on beamline ID29; statistics are given in Table 1. For the refinement of the structure, the better native data set was used (Nat20), but the Nat6 data set was used in the MIR phase calculations as it seemed more isomorphous to the Xe derivatives.
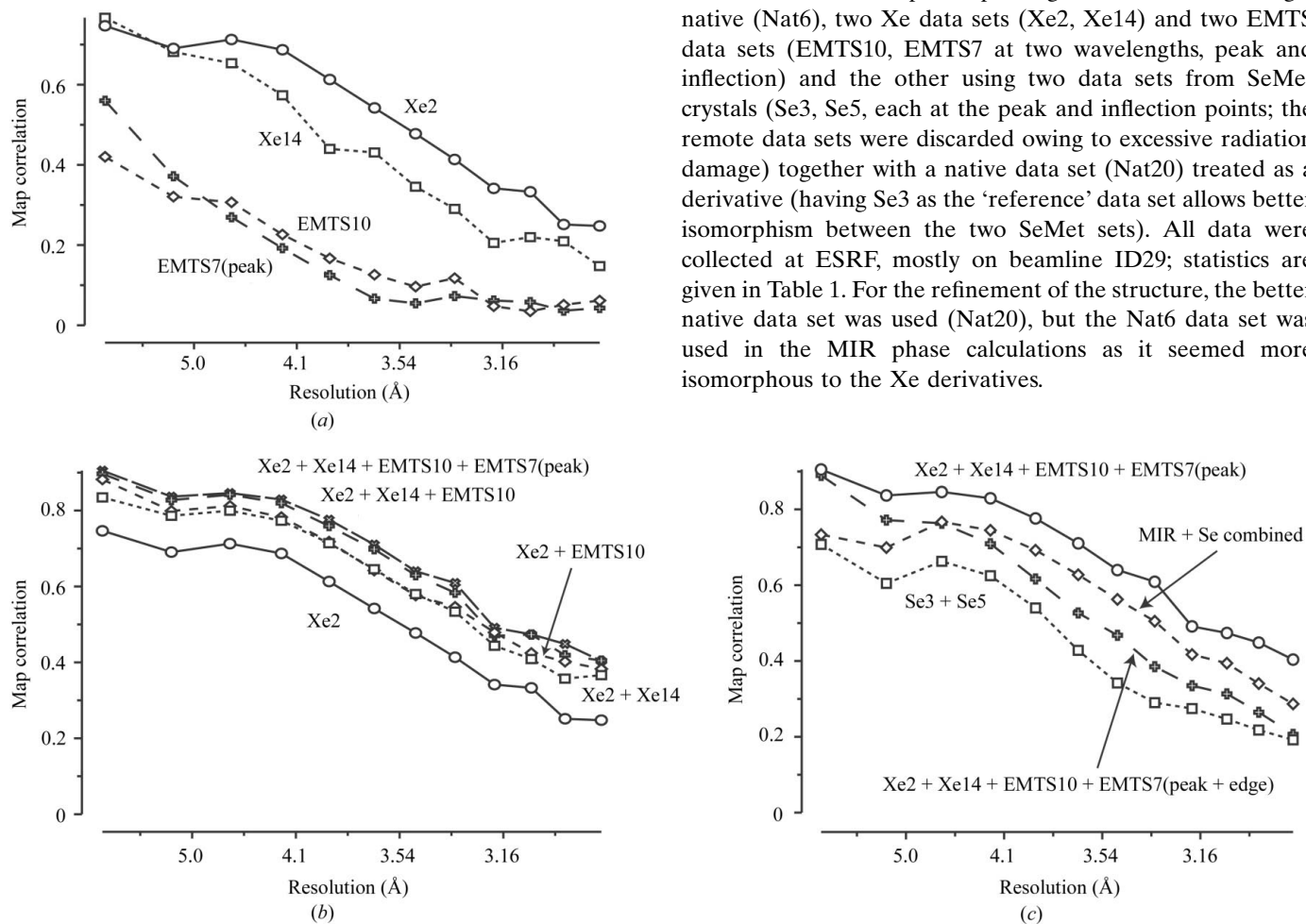
Figure 1
Correlations of observed and calculated structure factors as function of resolution for different experimental phase sets (equivalent to the correlation between maps). 'Observed' structure factors used the measured native amplitude (Nat6 or Se3) and the phase from *SHARP* and *SOLOMON* (solvent flattening with 52% solvent). (a) Single derivatives: the two Xe derivatives give much better phases than the EMTS derivatives. (b) Adding extra derivatives improves the phases: adding either Xe14 or EMTS10 to Xe2 gives about the same improvement, adding both is better still and adding EMTS7(peak) gives a further small improvement. (c) Phases are degraded by adding non-isomorphous derivatives: phases from the SeMet (peak and edge) data sets (together with native data set Nat20) are poor, and combining these phases with the best MIR set is less good than MIR alone. Adding the EMTS7(edge) data set also seriously degrades the MIR set.

2.1. Xenon

We had previously solved the structure of the C-terminal domain of the $\mu2$ subunit on its own using a Xe derivative (Owen & Evans, 1998), so we knew that Xe would bind to AP2. In order to enhance the anomalous signal, data were collected at long wavelength, Xe2 at 1.743 Å ($f'' = 9.0$) and Xe14 at 1.984 Å ($f'' = 11.0$). The optimum wavelength for Xe is a compromise between increasing the anomalous signal (up to a tabulated maximum of $f'' = 13.5$ at the L_I edge at 2.27 Å) and increasing errors arising from absorption at longer wavelength. The wavelengths of the Xe edges are too short or too long for routine data collection (K edge, 0.36 Å; L_I, 2.27 Å; L_{II}, 2.43 Å; L_{III}, 2.59 Å), but accessible wavelengths in the range 1.5–2.0 Å give excellent anomalous phasing. Cryoprotected crystals were pressurized to 1.0–1.2 MPa Xe for about 10 min, the pressure released and the crystals rapidly cooled to 100 K. The first major Xe site was found readily from the anomalous difference Patterson and extended to a total of 11 sites from residual maps from *SHARP*. One advantage of using two Xe-derivative crystals (produced at different pressures) is that the site occupancies are not the same, so they give different phase information.

2.2. Mercury

Two derivative data sets were used, both from crystals soaked for about 15 min in cryobuffer containing 1 m*M* EMTS. Crystal EMTS7 was used to collect data sets at three wavelengths, but the third remote-wavelength set was discarded because the radiation damage was too great. As will be seen below, the second (inflection or edge) set should have been discarded for the same reason, but was included in the phase calculations used to solve the structure.

2.3. Selenomethionine

The four subunits of the AP2 core complex were co-expressed in *Escherichia coli* from two plasmids in the presence of three different antibiotics (Collins *et al.*, 2002). Attempts to express the complex in the usual SeMet-supplemented minimal medium failed and it was necessary to add a 20% supplement of rich medium (LB) to obtain reasonable growth and expression. The incorporation of SeMet was thus less than 100% (incorporation level not measured), so the phasing power was less than might have been expected. Data from two crystals were used (Se3 and Se5): in each case, the third set from the remote wavelength was discarded owing to excessive radiation damage. The native data set Nat20 was included as a 'derivative' in these phase calculations, although it made little difference to the phasing. The phases from the SeMet derivatives were not particularly useful, but the sites were very useful in tracing the chain during the initial model building.

3. The quality of different phase sets

We can estimate how much each data set contributes to phasing and which is the best combination by comparing a series of phase calculations with single derivatives and combinations. Since we now know the structure, we can judge these by comparing the phase sets and maps with the refined model. Phase sets were compared by the complex correlation coefficient between the 'observed' structure factor (from the experimental phase calculation) $[m_{obs}|F_p|\exp(i\varphi_{obs})]$ and the calculated structure factor $[m_{calc}|F_p|\exp(i\varphi_{calc})]$ as a function of resolution. This is equivalent to a map correlation, but divided into resolution bins. It is not a perfect measure, since the refined model is far from perfect, but it should at least serve to rank the possible combinations of derivatives.

Fig. 1(*a*) shows the correlations for experimental phase sets for various single derivatives after solvent flattening [using *SOLOMON* (Abrahams & Leslie, 1996; Abrahams, 1997) run from the scripts in the *SHARP* interface, with the optimum 52% solvent]. This shows that the Xe derivatives are much better than the EMTS derivatives and that Xe2 is better than Xe14. Fig. 1(*b*) shows the cumulative effect of adding together derivatives. The two Xe derivatives together are better than either on their own: adding EMTS10 to Xe2 is slightly better than adding the second Xe, presumably because although weaker it is more different; the three derivatives Xe2, Xe14 and EMTS10 together are better than any pair and adding in the EMTS7(peak) data set makes a small improvement. This ranking of phase sets by map correlation is confirmed by visual inspection of the maps: adding EMTS10 to Xe2 is slightly better than adding Xe14 as the better low-resolution phasing improves connectivity.

The SeMet derivatives (Se3 and Se5) produced poorer phases (see Fig. 1*c*), as judged both by the map correlations and visual examination of maps. For the map used in the
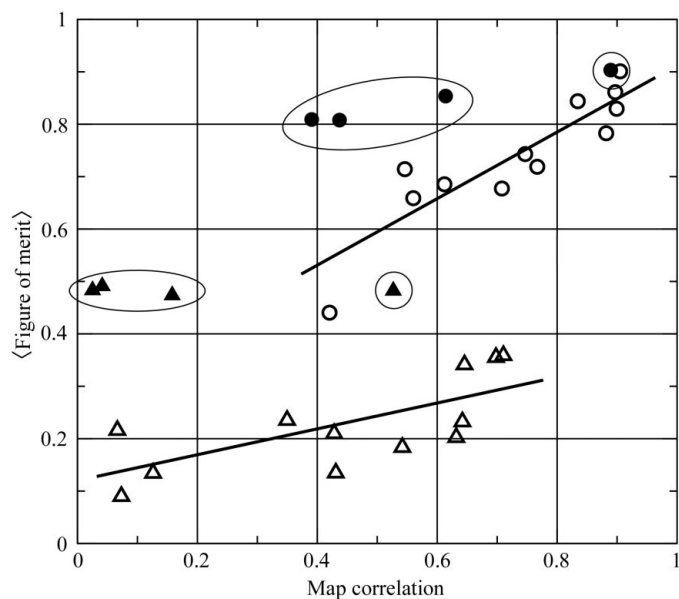
Figure 2
The mean figure of merit is generally well correlated with map quality, but phase sets including the poor EMTS7(edge) set have misleadingly high figures of merit (solid symbols, circled). Each symbol represents a particular phase set. Plotted for two resolution bins: triangles, ∞–9.7 Å resolution; circles, 4–3.65 Å resolution.

original build, two phase sets were combined: the MIR set [Nat6, Xe2, Xe14, EMTS10, EMTS7(peak and edge)] and the SeMet set (Se3, Se5, Nat20). The joint phases were improved by solvent flattening. The two phase sets were calculated separately because they need different 'reference' data sets for estimation of non-isomorphism: Nat6 for the MIR set and Se3 for the SeMet set. When this phase combination was first performed, the resulting map appeared to be better than either individually, but Fig. 1(c) shows that the joint phases are in fact less good than the MIR phases alone.

Originally, data at two wavelengths from EMTS7 were used: in retrospect, adding the second of these (edge) degraded the phases significantly, even in the presence of four other derivative data sets (Fig. 1c). The problem probably arises from radiation damage.

4. Assessment of derivatives

Phase correlations are of course no use in assessing derivatives or phase sets during the structure determination, but we do need measures which will tell us which derivatives to use and which phase set is best. Overall measures include the figure of merit (as a function of resolution), which measures strength and consistency of phase information, statistics from the solvent flattening, e.g. the correlation between E_{obs}^2 and E_{calc}^2 after the first cycle, and properties of the map such as the distribution of densities (Terwilliger & Berendzen, 1999a,b). Individual measures for each data set include phasing power and various R factors such as R_{cullis}. Do these measures allow us to choose the best phase set?

Of the overall measures, mean figure of merit does seem to be well correlated with map quality (Fig. 2), although some poor phase sets including the EMTS7(edge) data set have misleadingly high figures of merit (circled in Fig. 2). The solvent-flattening correlation coefficient does not seem to be a good indicator (not shown). Of the individual measures, both phasing power (Fig. 3) and R_{cullis} (not shown) seem reasonably reliable indicators of quality, although the EMTS7(edge) data set again has a misleadingly high isomorphous phasing power, even in the presence of other data sets. For the Xe derivatives, the anomalous phasing is dominant. Note that Xe14 has better phasing power statistics than Xe2 (Fig. 3), although its phasing
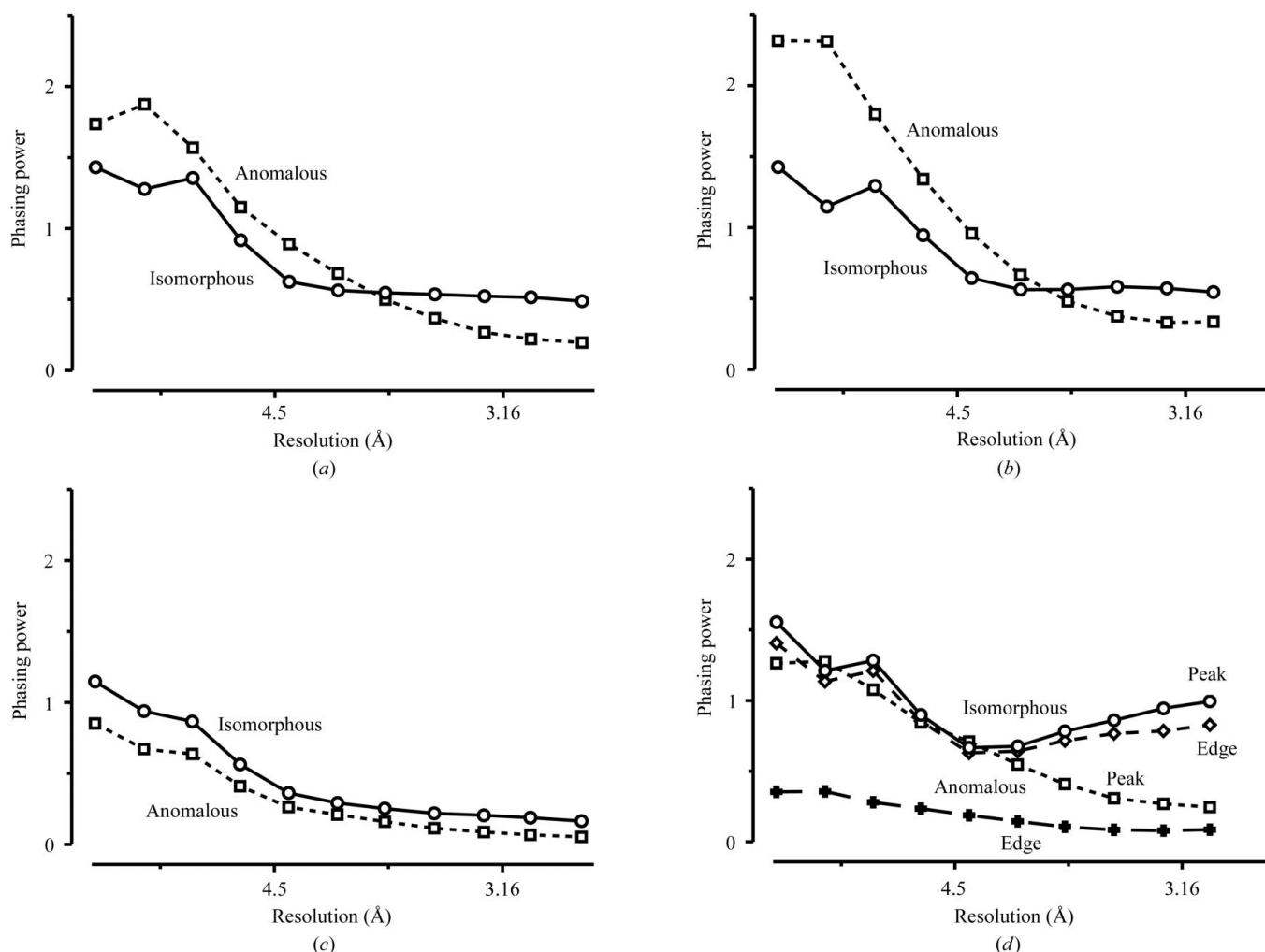
Figure 3
Phasing power *versus* resolution for (a) Xe2, (b) Xe14, (c) EMTS10 derivatives (from single-derivative phase sets) and (d) EMST7 (from the joint phase sets with all derivatives).
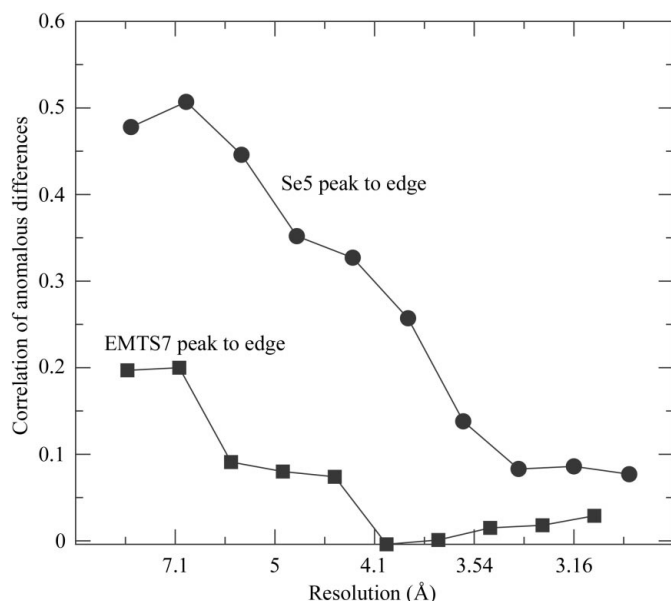
Figure 4
Correlation of anomalous differences ($|I_+| - |I_-|$) between peak and edge data sets for Se5 and EMTS7. The low correlation for EMTS7 indicates the unreliability of this data set.

is less good than Xe2, as judged by the map correlation statistic.

Could we have detected the unreliablility of the EMTS7(edge) data set? The best indicator of its dangers is the very poor correlation between the anomalous differences for the peak and edge data sets (Fig. 4): this is a clear hint that we should not have used this data set.

5. Conclusions

In cases such as this, where all the experimental phase information is weak, decisions on which data sets to use in phasing are not clear, nor is it always clear when sufficient information has been collected. In the end, the only reliable guide to the best map is interpretability, for either manual or automated model building. Interpretability depends on connectivity of density perhaps more than correctness in detail, so good and complete low-resolution phases are important. Adding more derivatives in MIR phasing is generally helpful, but adding in derivatives that are seriously non-isomorphous is detrimental. Radiation damage is a major cause of non-isomorphism.

No single statistic is a reliable indicator of effective phasing: the mean figure of merit is a reasonable guide, as are the individual phasing powers of each contribution, but all these indicators can be misleading. Where there are related pieces of phasing information, *e.g.* anomalous differences at different wavelengths, the correlation coefficient between observed differences is a good indicator of reliability. Existing statistics of phasing are not entirely reliable, which suggests that there is still potential for improvement of maximum-likelihood methods, despite their considerable success, for instance to provide better models of non-isomorphism and correlation between derivatives.

This postmortem investigation of various phasing strategies indicates that we did not make the best possible choice, but it was good enough to build the starting model from which we could refine the structure. It also confirms that Xe can provide good phases, even for large structures.

References

Abrahams, J. P. (1997). *Acta Cryst.* D**53**, 371–376.
Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.
Collins, B. M., McCoy, A. J., Kent, H. M., Evans, P. R. & Owen, D. J. (2002). *Cell*, **109**, 523–535.
La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.
Owen, D. J. & Evans, P. R. (1998). *Science*, **282**, 1327–1332.
Terwilliger, T. C. & Berendzen, J. (1999*a*). *Acta Cryst.* D**55**, 501–505.
Terwilliger, T. C. & Berendzen, J. (1999*b*). *Acta Cryst.* D**55**, 1872–1877.