Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Faster data-collection strategies for structure determination using anomalous dispersion

Ana González

Stanford Synchrotron Radiation Laboratory, 2575 Sand Hill Road, MS99, Menlo Park, CA 94025, USA

Correspondence e-mail: ana@slac.stanford.edu

Received 23 September 2002
Accepted 29 November 2002

Many macromolecular structures are being determined using anomalous dispersion phasing methods. Different data-collection strategies at one, two, three or more wavelengths can be used for these experiments. The choice of strategy can determine the success or failure of the experiment and should be based on a clear understanding of the advantages and disadvantages of each approach given the experimental constraints and goals. In this paper, several sets of three-wavelength MAD experiment data were reanalyzed using one, two and three wavelengths and systematically removing reflections from the data sets to determine the minimum amount of data required to yield an automatically traceable map as a function of the number of wavelengths used in phasing. In the cases studied here, two-wavelength MAD consistently required fewer data than three-wavelength MAD, as long as the unique data completeness was high at each wavelength. It was also found in some instances that using one wavelength for phasing required as much or more data as using two wavelengths. These results can help with the design of adequate data-collection strategies which maximize the phasing power from the minimal data collected. This is particularly important for minimizing the effects of radiation damage on phasing while taking sample characteristics, beamline properties and experimental goals into account.

1. Introduction

De novo crystal structure determination using MAD methods is one of the key steps in both structural genomics projects and conventional structural biology (Smith *et al.*, 1996; Hendrickson, 1999; Ealick, 2000). The ease of the experiment in modern dedicated beamlines (Roth *et al.*, 2002; Pohl *et al.*, 2001) and the use of cryogenic techniques to increase the lifetime of the sample (Garman, 1999) have played a large role in the success of these experiments.

A perusal of publications of structures solved by MAD (see, for example, the compilation by Hendrickson & Ogata, 1997) shows that a very common data-collection strategy consists of collecting data at three wavelengths or sometimes four. Inverse-beam geometry is often used to collect Friedel related reflections. This strategy is devised to obtain highly accurate experimental phases, because the choice of wavelengths is such that both the anomalous and dispersive contributions to the phasing are optimized and a very redundant set of anomalous (measured from acentric pairs, *i.e.* Friedel or Bijvoet related reflections) and dispersive differences (measured between the same reflections at different wavelengths) are obtained. This data-collection strategy requires

six times (eight times with four wavelengths) the amount of data required for a standard complete data set to the same resolution (Dauter, 1997).

Sometimes the acentric pairs are collected in a single angular segment instead of using inverse-beam geometry. Although in general this strategy does not cancel systematic errors between acentric pair measurements,[1] these errors can be corrected to a large extent during scaling (Hendrickson & Teeter, 1981; Friedman *et al.*, 1995; Evans, 1997). In many cases, this strategy can reduce the data needed by as much as half, but for some crystal symmetries and orientations the same amount of data as for the inverse-beam strategy would still be required (Dauter, 1997).

Long experiments can be a problem because the longer the experiment lasts, the higher the radiation dose the crystal will receive, increasing the risk of specific structural changes in the sample caused by radiation damage (Burmeister, 2000; Leiros *et al.*, 2001). Collecting data at all wavelengths simultaneously in small angular wedges, as suggested by Rice *et al.* (2000), would be the optimal way to minimize the effect of a long data collection on the phases. If the crystal is extremely sensitive to radiation, this strategy could result in three incomplete data sets which cannot be used for phasing unless more data are collected from another crystal or, if the crystal is much larger than the beam in the spindle direction, from an unexposed part of the crystal.

Alternatives to further shorten the time required for the experiment have been suggested. One is to carry out a two-wavelength MAD data collection (Okaya & Pepinsky, 1956). This has been shown to be feasible in several cases provided that one of the two wavelengths has a small f' and relatively large f''. This condition is often achieved with a remote wavelength on the high-energy side of the absorption edge (Peterson *et al.*, 1996; González *et al.*, 1999). Another option suggested is SAD data collection. For instance, Dauter *et al.* (2002) and Rice *et al.* (2000) studied a large number of cases and proved that single-wavelength phasing (Wang, 1985) can be successful regardless of diffraction resolution, anomalous scatterer and anomalous signal. An advantage of single-wavelength methods over two or more wavelength MAD is versatility: while MAD experiments can only be properly performed at tunable synchrotron beamlines fulfilling certain requirements for wavelength bandpass, stability and reproducibility (Thompson, 1997), SAD data can be collected at any macromolecular crystallography beamline. Even the Cu $K\alpha$ emission at home sources can be used to successfully phase structures (Jaskólski & Wlodawer, 1996; Dauter *et al.*, 1999; Yang & Pflugrath, 2001).

Under current experimental conditions at most beamlines, a case can be made for collecting as much MAD data as possible while the crystal lasts or, if data processing can keep up with data collection, until the user has analyzed the data and produced a clearly interpretable map. This favors the longer three- or four-wavelength data-collection strategies described above. The development of data-collection and processing facilities for high-throughput projects (Kuhn & Soltis, 2001; Abola *et al.*, 2000; Ferrer, 2001) could have an influence on the choice of strategy: an increasing proportion of experiments will be carried out semi- or fully automatically and it is likely that samples belonging to different projects will be stored together for data collection, with the possibility of remounting a sample easily if the data collected previously proved to be insufficient for structure solution without having to reschedule beam time for the project. Under these circumstances, it is important to collect data in the most time-efficient manner, using strategies which minimize the total experiment time without compromising the quality of the data.

In the case of experiments that are being conducted at rapidly tunable beamlines, a careful analysis is required to determine the actual minimal amount of data required to successfully phase a structure. The impact on phasing information from multiple-wavelength data *versus* completeness and redundancy is directly related to the amount of data and hence the time required for each of the approaches listed above. The purpose of the study presented here is to explore the minimum data amount needed to solve a structure with typical quality data with one-, two- and three-wavelength phasing, and to propose possible strategies for automated data collection.

2. Methods

A total of six samples were selected for this study. Of those samples, five were selenomethionine proteins. Selenium is the most commonly used anomalous scatterer in macromolecular MAD experiments because of the well developed method of substituting the sulfur in methionine by selenium (Doublié, 1997) and it is the anomalous scatterer of choice for high-throughput projects (Lesley *et al.*, 2002). All the selenium-containing samples were produced by the Joint Center for Structural Genomics (JCSG) project to obtain the structure of the proteins coded by the *Thermotoga maritima* bacterium genome (Lesley *et al.*, 2002). The other sample was recombinant sperm whale myoglobin (Mb), an iron metalloprotein. Some relevant sample characteristics are listed in Table 1. A wide range of crystal symmetries, solvent content and anomalous signal are represented by the samples. All those factors may play a role in the amount of data needed to solve the structure.

2.1. Data collection

Data collection for all crystals was performed at the fully tunable dedicated SSRL MAD beamline 9-2. Data at three wavelengths were collected for each crystal. For the Mb and Tm1084 crystals, the MAD data used were collected solely for the purpose of this study. In the other cases, the data used are the same data from which the corresponding structures were originally solved.

For all the samples, data collection was performed at the 'peak' (maximum f'') and inflection point (minimum f') of the

[1] Except in the special case where the crystal is set with an even-fold symmetry rotation axis perpendicular to the spindle.

Table 1
Sample characteristics and results of the data analysis using all data collected.

The merging statistics are given for the remote-wavelength data set.

	Mb	Tm0423	Tm0665	Tm1083	Tm1102	Tm0667
PDB code	1jw8	1kq3	1j6n	1j5u	1o0w	1j6o
No. anomalous scatterers†	1 (1)	7 (7)	48 (46)	1 (0)	10 (8)	1 (0)
Fractional anomalous signal‡	0.03	0.06	0.09	0.03*	0.06	0.02*
No. of residues§	150	364	1164	124	480	256
Solvent content¶ (%)	55	40	40	60	43	29
Space group	$P6$	$I422$	$P2_1$	$P4_332$	$P1$	$P2_12_12_1$
Oscillation†† (°)	270	52.5 × 2	96 × 2	45 × 2	360	120 × 2
Maximum resolution (Å)	2.1	2.0	2.0	2.1	2.0	2.0
Resolution limit of anomalous signal‡‡ (Å)	2.1	2.0	2.1	3.2	2.7	2.7
$I/\sigma(I)$§§	6.5 (2.6)	11.7 (6.5)	11.4 (3.6)	7.6 (1.6)	6.8 (2.5)	5.7 (2.3)
$R_{\rm meas}$§§¶¶	0.11 (0.34)	0.06 (0.12)	0.04 (0.15)	0.11 (0.55)	0.09 (0.38)	0.1 (0.33)
Completeness††† (%)	100	100	87.5	100	93.7	100
Acentric completeness††† (%)	100	100	87	100	92.9	100
Multiplicity	14.8	8.1	4.1	17.9	3.9	9.1
Residues traced (%)	99	97	88	95	95	86

† Values in parentheses are the number of sites with a temperature factor higher than 50 Å^3. ‡ Defined as $(N/2)^{1/2}2f''/|F|$, where N is the number of ordered anomalous scatterer sites in the asymmetric unit and $|F|$ is the total structure-factor amplitude (in the cases labelled *, the single disordered anomalous scatterer was assumed to have an occupancy of 0.5). § Total number of residues in the asymmetric unit. ¶ Estimated from the contents of the unit cell. †† Total φ range collected at each wavelength (factored by 2 when an inverse-beam pass was performed). ‡‡ Resolution to which the correlation of the anomalous differences is larger than 0.3. §§ Values in parentheses are for the highest resolution bin. ¶¶ R factor weighted for multiplicity (Diederichs & Karplus, 1997). ††† Unique and acentric data completeness for one wavelength.

corresponding absorption edge and at a remote wavelength. For the crystals containing selenium, the inflection, peak and remote wavelengths were around 0.979, 0.98 and 0.92 Å, respectively. The data collection on Mb was performed at the peak and inflection points of the iron K absorption edge (1.738 and 1.739 Å) and at a remote wavelength of 1.36 Å.

A suitable oscillation starting point to optimize data completeness was selected for each MAD experiment, depending on the crystal orientation and space group. The 'strategy' option in *MOSFLM* (Leslie, 1991) was used to choose the appropriate data-collection strategy. All the crystals were in random orientation, *i.e.* no effort was made to align a crystal symmetry reciprocal axis with the oscillation axis. Redundant data were collected for all cases either using inverse-beam geometry or, for Mb and Tm1102, in a single segment pass. All the data had consistent scaling statistics throughout the whole experiment, except for four frames in the Tm0665 data at the maximum f'' wavelength, which were removed from the data set. This indicates that the crystals did not decay substantially because of radiation damage.

2.2. Data analysis

Data were processed with the program *MOSFLM* (Leslie, 1991) and scaled and merged with *SCALA* (Evans, 1997). The amplitudes were calculated with *TRUNCATE* (French & Wilson, 1978). All these programs are part of the *CCP*4 program suite (Collaborative Computational Project, Number 4, 1994). Phasing and density modification were carried out with *SOLVE* and *RESOLVE* (Terwilliger & Berendzen, 1997; Terwilliger, 2000). The quality of the maps before and after density modification was assessed by means of the correlation coefficient to the map calculated from a refined model. The program *OVERLAPMAP* (Collaborative Computational Project, Number 4, 1994) was used to calculate the correlation coefficient.

Tracing of the main chain was attempted with the automatic procedure implemented in *ARP/wARP* v. 5.1 and v. 6.0 (Perrakis *et al.*, 1997; Morris *et al.*, 2002). The tracing of the model was considered to be successful if the connectivity value given by the autotracing program was higher than 0.9. In terms of the structure, this corresponded to between 88 and 98% of the residues being traced, depending on the sample. This criteria provides a straightforward and objective way to compare the results of different phasing scenarios for the same crystal structure. However, it might lead to an underestimation of the success of the phasing procedure, because many low-resolution maps that cannot be fully autotraced may be still be interpretable. To avoid the problem of subjective judging of the map quality, only samples diffracting to around 2 Å were chosen, because current autotracing algorithms work most reliably with data to this or better resolution. The qualitative results are expected to hold at most common maximum diffraction resolutions. However, alternative phasing methods could be considered in the case of very high resolution data, such as direct methods (Hauptman, 1997; Foadi *et al.*, 2000) or phasing from a single atom position (Benini *et al.*, 2000), which, whether in combination with anomalous dispersion or not, could result in more time efficient experiments than using anomalous dispersion methods alone.

SAD phasing was performed with the data collected at the peak wavelength. The two-wavelength MAD phasing was performed using the remote and inflection wavelengths, since this wavelength combination has been found to give the best phases in the two-wavelength experiments (González *et al.*, 1999). The results from the processing and phasing using all data collected for each crystal are shown in Table 1.

The scaling and phasing procedure described above was repeated for each sample and for each phasing scheme (using one, two or three wavelengths), each time removing sequential frames from the data set starting with the inverse-beam pass. Depending on the crystal symmetry, 10 to 2° of data were removed at each time, until the chain could no longer be autotraced. This systematic way of removing reflections from the data set used for phasing reproduces the real-life situation in which part of the diffraction data is unusable because of radiation damage or the data collection is terminated because of time constraints or other restrictions, assuming that the data collection is performed in small-angle wedges and the inverse-

Table 2
Minimal data sets required to solve the structure automatically by SAD and two- and three-wavelength MAD.

For MAD phasing, the unique data, acentric completeness and multiplicity refer to the inflection-wavelength data set. The correlation coefficient to the refined model is given both for the map calculated after density modification (DM) and the experimental map (Exp).

		Completeness (%)			Correlation (%)		
Sample	φ† (°)	Unique	Acentric	Multiplicity	DM	Exp	Relative dose‡
Three-wavelength MAD							
Mb	135	92	44	2.7	66	36	0.05
Tm0423	72	88	60	2.0	65	44	0.04
Tm1083	30	91	70	2.7	66	35	0.03
Tm0665	288	86	66	2.1	62	45	0.11
Tm1102	750	93	77	2.0	50	35	0.80
Tm0667	540	100	100	7.9	50	36	0.32
Two-wavelength MAD							
Mb	110	96	64	3.3	67	37	0.04
Tm0423	51	91	64	2.0	66	43	0.02
Tm1083	20	91	70	2.7	65	35	0.02
Tm0665§	326	86	74	3.5	67	45	0.09
Tm1102	700	94	93	3.7	51	35	0.60
Tm0667	360	100	100	7.9	51	35	0.18
SAD							
Mb	170	100	100	9.4	61	32	0.05
Tm0423	52.5	100	97	4.1	60	37	0.04
Tm1083	20	99	97	4.7	62	27	0.03

† φ is the total angular range of data necessary to solve the structure, *i.e.* the oscillation range collected at each wavelength multiplied by the number of wavelengths used. ‡ Estimate of the dose absorbed by the crystal for the minimal data set, in fractions of the Henderson limit (2×10^7 Gy; Henderson, 1990). § The figures given for this case correspond to the instance where all the automatically found Se sites were included in the phasing. When no incorrect sites were used (this could be achieved by rejecting the ten sites with lowest occupancy), the structure could be solved with a total φ rotation range of 222°.

beam segment is collected after a complete data set at each wavelength has been secured. This strategy has been suggested by Rice *et al.* (2000) as a safe way to collect MAD data in order to forestall the effects of radiation damage on the phasing. The data set containing the minimum amount of data to solve the structure according to the above criteria with single or multiwavelength phasing will from now on be referred to as the 'minimal' data set.

An estimate of the dose absorbed by the crystal during collection of the minimal data set was made based on the theoretical values for the intensity of the beamline and mass absorption coefficients at the wavelengths used for data collection. A regular cubic crystal shape and uniform illumination by the X-ray beam were also assumed for the calculation. This is likely to result in an overestimation of the values for the absorbed dose.

3. Results

All the structures in the study could be solved as described above using either three or two wavelengths, with fewer data than the total collected. Half of the structures could also be traced after SAD phasing. In the other cases this was not possible even using all data collected at the peak wavelength. The models obtained after tracing had the same or a similar number of residues to those traced using all data. Table 2 summarizes the results of phasing with the minimal data. The discussion below will focus mainly on the results obtained with the minimal data sets, unless stated otherwise.

3.1. Location of anomalous scatterer sites

Although determining the minimum amount of data to be able to locate the heavy-atom sites was not the aim of the study, it was confirmed that solving the anomalous scatterer substructure is not the critical factor determining the size of the minimal data set. In all cases but one, the correct sites were located with fewer data than needed to solve the crystal structure. The exception was the Tm0665 example, with 48 Se sites in the asymmetric unit, where only six sites could be found with the SAD data. Although most sites were located using two-wavelength data, the inclusion of some incorrect sites also increased the size of the minimal data set for two-wavelength MAD phasing. When care was taken that only correct sites were used (by rejecting lower occupancy sites), it was possible to arrive at the structure with substantially fewer data (see Table 2). The structure could also be solved by SAD phasing with the data available, provided that the correct sites were supplied. This example raises the question of whether finding the anomalous scatterer sites might be the critical step determining the amount of data needed to solve the structure in other similar cases with a large number of anomalous scatterers.

3.2. SAD and MAD phasing

The results show that using the procedure and criteria described above, phasing with just one wavelength required complete data sets including measurements of all or almost all anomalous pairs. SAD phasing with Tm1083 and Tm0423 data required about the minimum oscillation to achieve complete data set in the respective crystal space groups and orientations. In the case of Mb, a larger amount of redundancy was needed. In this example, the crystal orientation was such that a nearly full set of Friedel related pairs could only be collected with about 180° of data. In addition, the Mb peak data may have suffered from larger systematic absorption errors resulting from data collection at long wavelengths.

The structures with the lowest symmetry (Tm1102, Tm0665 and Tm0667, see Table 1) could not be autotraced with the data collected at one wavelength. As mentioned above, in the case of Tm0665 the heavy-atom substructure could not be solved using single-wavelength data, which precluded subsequent phasing. The peak data sets for this case and Tm1102 were only 94 and 87% complete, respectively, with approximately fourfold redundancy, as shown in Table 1. Tm0667 had an extremely low anomalous signal, with a single N-terminal selenomethionine in 256 residues. The need for large data multiplicity for SAD phasing when the anomalous signal is small has been illustrated by Dauter & Adamiak (2001). It is most likely that SAD phasing would have worked in all these three cases if more data had been available.

3.3. Two- and three-wavelength MAD phasing

Phasing with either two or three wavelengths led to a traceable map with less than full data completeness and many missing acentric pairs in all but one example. In three of the cases studied (Mb, Tm0665 and Tm1102), using data at two wavelengths required better completeness than using three wavelengths, both in the total unique reflections and acentric reflections. The total amount of data needed was always smaller when using two wavelengths. For the other examples, approximately the same amount of data was needed at each wavelength for two- and three-wavelength MAD. The map correlation to the model-based map, shown in Table 2, is roughly the same whether using two or three wavelengths. For Tm0665, the correlation of the two-wavelength map after density modification is significantly better. In this particular case, the two-wavelength minimal data had much higher redundancy than the three-wavelength data set.

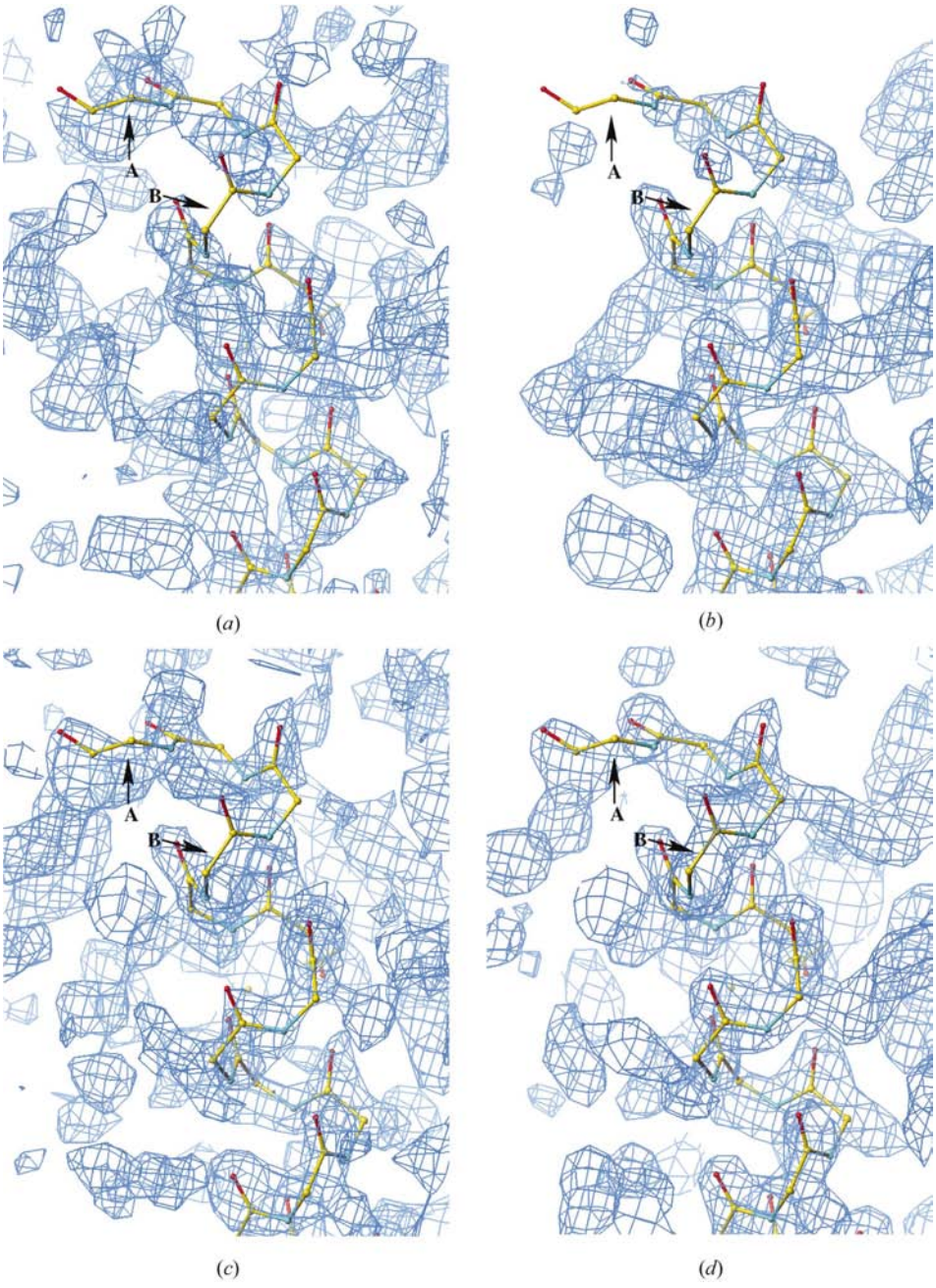

Figure 1
Tm0423 maps calculated from SAD and MAD phases using the same total amount of data (52.5°). (*a*) SAD experimental map. (*b*) SAD map after density modification. (*c*) Two-wavelength MAD experimental map. (*d*) Two-wavelength MAD after density modification. The part of the structure shown is an α-helix near the surface of the protein (the helix backbone is displayed). The arrows near the top of the view labeled *A* and *B* point to zones where the density-modified SAD map (*b*) has breaks in the electron density. The experimental SAD map (*a*) shows density for the *A* zone, but this density disappears after density modification. The two-wavelength MAD maps (*c*) and (*d*) exhibit continuous density over these areas.

3.4. Map quality

The experimental MAD maps calculated with the minimal data sets are clearly better than the SAD maps (see Fig. 1). This is expected because SAD phasing results in bimodal phase distributions for all the reflections. After density modification, the SAD maps show a proportionally larger improvement than the MAD maps. This has also been observed by Rice *et al.* (2000). The SAD maps always had a slightly lower correlation and tended to lack connectivity compared with the MAD maps obtained with the same amount of data. The two- and three-wavelength maps are of similar quality and are equally improved by density modification.

Before autotracing, the map correlation is somewhat lower when phasing with the minimal amount of data than when the complete data sets with higher reflection multiplicity are used. As shown in Table 2, when phasing with the minimal data sets the map correlation coefficients are between 0.3 and 0.4 for the experimental maps and are 0.5–0.7 for the density-modified maps in most cases. When phasing with the data sets containing all the data, the typical experimental correlation coefficients rise to between 0.4 and 0.6 and, after density modification, are between 0.5 and 0.8. After autotracing, a very similar model was always arrived at independently of the number of wavelengths and the completeness of the data

employed for phasing. More autotracing cycles were needed to obtain a complete model as the data completeness decreased.

4. Discussion

A critical factor determining the minimum amount of data necessary to solve the structure is the number of reflections with a sufficiently well determined phase to serve as an adequate starting point for solvent-flattening and phase-extension schemes to work successfully. With 'perfect' phases calculated from a model, structure solution has sometimes been found to succeed with about only 60% of the unique reflections. When the phases must be calculated from the small anomalous and dispersive differences derived from the experimental data, the total number of measurements needed to achieve the critical number of phased reflections is obviously much larger.

A comparison of the results of single- and multiple-wavelength phasing for the same structure indicates that multiple-wavelength phasing consistently requires less data completeness and redundancy than the single-wavelength counterpart. A likely explanation is that both the dispersive and anomalous differences contribute to the phasing in the former case. In a standard MAD experiment, with the same zone of the crystal sampled at each wavelength, the number of dispersive difference measurements will be roughly equal to the number of unique reflections, regardless of space group and crystal orientation. In contrast, a reasonably complete set of anomalous difference measurements, which are the only source of experimental phase information in SAD phasing, is not in general achieved with a complete set of unique reflections (Dauter, 1997).

Another contribution to the number of phased reflections in MAD experiments, which may be of some importance in some space groups, is made by centric reflections. These do not contribute to experimental SAD phases but, because the dispersive term f' always causes a change in the amplitude of the scattering factor at different wavelengths, are phased in MAD experiments.

Two-wavelength MAD can in some cases require more data at each wavelength than three-wavelength MAD because the value of f'' is relatively small at the inflection and remote wavelengths used for the two-wavelength phasing analysis, which can result in somewhat fewer accurate phases than when both anomalous and dispersive differences are optimized.[2] Other sample-dependent factors, such as good diffraction quality and high anomalous signal, are likely to downweight the contribution of a third wavelength.

4.1. Effect of sample properties

Intrinsic structure properties are relevant to the success of phasing. The crystal symmetry directly influences the redundancy of the data. In general, the higher the symmetry, the easier it will be to solve the structure with fewer data than the standard rotation angle for the space group in question. This is illustrated very well by the examples presented here.

A high anomalous-to-elastic scattering ratio results in large anomalous and dispersive differences which are easier to measure despite some errors in the data, although a very large number of anomalous scatterers can offset the advantage of a high anomalous signal because solving the heavy-atom substructure becomes more difficult, as seen for Tm0665. In cases such as this, prior knowledge of the heavy-atom positions (for example, from isomorphous differences) could shorten the experiment and facilitate automatic structure solution and should be taken into account when deciding on the data-collection strategy.

High data quality is also important. In cases where the crystals diffract poorly, choosing an adequate exposure time to measure experimental intensities accurately can decrease the need for data redundancy.

Favorable crystal properties might be comparatively more critical when attempting SAD phasing because, as discussed above, anomalous differences are more sensitive to systematic errors of the data.

4.2. Density modification and autotracing

In the present study, the SAD maps were slightly but consistently worse than the MAD maps after density modification, even though the minimal MAD data sets were more incomplete. This may mean that the better quality of the MAD experimental phases is the critical factor in the performance of density modification, rather than the better completeness and redundancy of the SAD minimal data set. It is possible that some of the quality differences between MAD and SAD maps could be attributed to 'over-solvent-flattening' of disordered areas of the structure, as shown in Fig. 1. This can generate disconnected areas in the maps, particularly at the protein–solvent boundary, and make autotracing harder. It could be advantageous to use direct methods to solve the phase ambiguity in SAD phasing (Fan & Gu, 1985; Fan *et al.*, 1990; Langs *et al.*, 1999).

The results also suggest that thanks to the phase information available from even a partial model, autotracing is more powerful than other density-modification procedures, since the resultant model completeness does not on depend on the number of wavelengths or the exact amount of data used for phasing from the point where a minimally interpretable map is available. This means that a large number of experiments seeking to solve unknown structures would not be affected by the slight loss of accuracy of the experimental phases obtained with lower redundancy data sets.

5. Data-collection strategy

The comparison of the two- and three-wavelength MAD phasing shows that two-wavelength MAD is a more time-effective experiment, at no significant cost in the model

[2] Note that optimizing $\Delta f'$ rather than f'' remains the optimal strategy for two-wavelength MAD phasing (Peterson *et al.*, 1996; Hädener *et al.*, 1999; González *et al.*, 1999), probably because dispersive differences can be measured more accurately in a standard MAD experiment. For a full discussion, see González *et al.* (1999).

quality. The results obtained for SAD phasing with Mb, Tm1843 and Tm0423 appear to indicate that a two-wavelength MAD experiment would not as a rule require more data than SAD experiments and can even be shorter in some cases. Therefore, two-wavelength MAD appears to be as suitable as SAD to increase the productivity of dedicated macromolecular crystallography beamlines.

Regarding the decrease of the total radiation dose absorbed by the crystal, a two-wavelength MAD data collection at the inflection and remote wavelengths would actually be better than SAD, since the maximum f'' wavelength optimal for SAD experiments is that where most radiation is absorbed. This hypothesis is based on the rough estimation of the dose absorbed by the crystal during the exposure time required to collect the minimal data set (Table 2); this has yet to be confirmed experimentally. In addition, the MAD maps also tend to be slightly better and, as suggested by the Tm0665 example, using more than one wavelength can be useful to determine the anomalous scatterer substructure in difficult cases.

Taking these points into account, the optimal strategy for MAD data collection in a dedicated beamline with automatic sample mounting would be to collect data at the inflection of the absorption edge and a remote wavelength far away from the edge. A continuous oscillation range providing close to 95% unique completeness would work well for the majority of cases. Because Friedel pair completeness does not appear to be critical, collecting data in a continuous oscillation segment rather than using the inverse-beam geometry is preferable. Even in the worst-case scenario that the crystal was damaged before the data collection is complete, this strategy would minimize the number of crystals needed to solve the structure.

If the MAD data were also to be used for structure refinement, it would be convenient to aim for close to 100% completeness at least at one wavelength. This can be achieved with very little extra exposure by collecting a non-contiguous rotation segment, particularly if the crystal is mis-set for this purpose.

With automated sample mounting, the sample could be retrieved easily in cases when an unfavorable anomalous-to-total scattering ratio or poor-quality diffraction made it necessary to collect more data (assuming that the data analysis cannot keep pace with the data collection). Alternatively, if the data-collection software was able to access information about initial screening results and sample characteristics from a database, as is the case for some structural genomics projects, an inverse-beam or a similar data-collection strategy aimed at collecting redundant data could be programmed right at the start for potentially difficult experiments such as Tm0667. In cases such as this, *a priori* knowledge of the very small anomalous signal, low solvent content and moderate diffraction quality makes it easy to predict that more than the minimum standard rotation for the space group is likely to be needed to solve the structure. Extra redundancy can also be useful in experiments aiming towards obtaining an undistorted view of the solvent area or fine structural detail. For these experiments, more accurate experimental MAD phases can be helpful in interpreting the ordered solvent structure (Burling *et al.*, 1996; Schmidt *et al.*, 2002) or guiding the structure refinement (Coste *et al.*, 2002). Using direct methods in combination with MAD (Gu *et al.*, 2001) might help shorten the experiment in these cases.

Because in the general case f'' is not optimized at the wavelengths suggested for two-wavelength experiments, collecting data simultaneously at both wavelengths in small φ wedges would be preferred to collecting one wavelength at a time. The former strategy would be better at preserving the dispersive differences in case of damage to the sample and to facilitate data analysis on the fly, deriving the heavy-atom substructure from the dispersive differences alone or combined with the anomalous differences. If automatic data analysis is not implemented and it is possible to evaluate the data as they are being collected, as is the case with less intense X-ray sources, it may be more practical to collect one entire wavelength at a time, rather than in wedges, to simplify data processing. In this case, collecting the wavelength with the highest f'' first would allow the fastest determination of the heavy-atom substructure and give the best chance of solving the structure by SAD in case the data collection could not be finished.

Redundant SAD data collection would be a better option than MAD when instrumentation constraints make it difficult or impossible to collect a remote wavelength sufficiently far away from the absorption edge, when the beamline stability, bandpass or reproducibility are not appropriate for MAD and, of course, if the absorption edge of interest is not available.

6. Summary

The examples presented in this paper, representing typical data quality and resolution and processed by conventional methods, show that both two-wavelength MAD at the inflection and remote wavelength, and SAD at the peak wavelength can result in interpretable maps using significantly fewer data than three-wavelength MAD. It was also found that for the samples used in the study, SAD phasing did not require fewer data than two-wavelength MAD phasing. Therefore, both single or two-wavelength experiments may be equally suitable for maximizing beamline throughput and decreasing the exposure of the crystal to radiation. When fulfillment of these requisites is crucial for the experiment, as is the case for high-throughput projects, in experiments with radiation-sensitive samples, when the available beam time is limited *etc.*, the choice of strategy should be made based on the capabilities of the X-ray source and the characteristics of the sample.

In addition to the obvious case where the relevant absorption edge is not accessible at a dedicated beamline, collection at a single wavelength may be advantageous when the beamline wavelength range is limited and a suitable remote wavelength far away from the absorption edge cannot be reached. On the other hand, when the beamline is suitable for MAD and tunable over a large range about the absorption edge of interest so that a large $\Delta f'$ value can be achieved, two-wavelength MAD would be the more advantageous strategy,

both because of the somewhat higher accuracy of the phases, which provides an advantage at the model refinement stage (Murshudov *et al*., 1997), and the possibility of decreasing the radiation damage to the crystal by avoiding collection at the maximum f'' wavelength. A controlled experiment to assess the damage suffered by the crystal under different data-collection strategies would be worthwhile.

Much of the data used for this study were collected and processed by the Joint Center for Structural Genomics staff at the SSRL. The Stanford Synchrotron Radiation Laboratory is operated by Stanford University on behalf of the US Department of Energy, Office of Basic Energy Sciences. The SSRL Structural Molecular Biology Program is supported by the Department of Energy, Office of Biological and Environmental Research and by the National Institutes of Health, National Center for Research Resources, Biomedical Technology Program and the National Institute of General Medical Sciences.

References

Abola, E., Kuhn, P., Earnest, T. & Stevens, R. C. (2000). *Nature Struct. Biol.* **7**, *Suppl.*, 973–977.
Benini, S., González, A., Rypniewski, W. R., Wilson, K. S., Van Beeumen, J. & Ciurli, S. (2000). *Biochemistry*, **39**, 13115–13126.
Burling, F. T., Weis, W. I., Flaherty, K. M. & Brünger, A. T. (1996). *Science*, **271**, 72–77.
Burmeister, W. P. (2000). *Acta Cryst.* D**56**, 328–341.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Coste, F., Shepard, W. & Zelwer, C. (2002). *Acta Cryst.* D**58**, 431–440.
Dauter, Z. (1997). *Methods Enzymol.* **276**, 326–344.
Dauter, Z. & Adamiak, D. A. (2001). *Acta Cryst.* D**57**, 990–995.
Dauter, Z., Dauter, M., de La Fortelle, E., Bricogne, G. & Sheldrick, G. M. (1999). *J. Mol. Biol.* **289**, 83–92.
Dauter, Z., Dauter, M. & Dodson, E. (2002). *Acta Cryst.* D**58**, 494–506.
Diederichs, K. & Karplus, P. A. (1997). *Nature Struct. Biol.* **4**, 269–275.
Doublié, S. (1997). *Methods Enzymol.* **276**, 326–344.
Ealick, S. E. (2000). *Curr. Opin. Chem. Biol.* **4**, 495–499.
Evans, P. R. (1997). *Proceedings of the CCP4 Study Weekend. Recent Advances in Phasing*, edited by K. S. Wilson, G. Davies, A. W. Ashton & S. Bailey, pp. 97–102. Warrington: Daresbury Laboratory.
Ferrer, J.-L. (2001). *Acta Cryst.* D**57**, 1752–1753.
Fan, H. F. & Gu, Y. X. (1985). *Acta Cryst.* A**41**, 280–284.
Fan, H. F., Hao, Q., Gu, Y. X., Qian, J. Z., Zheng, C. D. & Ke, H. M. (1990). *Acta Cryst.* A**46**, 935–939.
Foadi, J., Woolfson, M. M., Dodson, E. J., Wilson, K. S., Jia-xing, Y. & Chao-de, Z. (2000). *Acta Cryst.* D**56**, 1137–1147.
French, G. S. & Wilson, K. S. (1978). *Acta Cryst.* A**34**, 517.
Friedman, A. M., Fischman, T. O. & Steitz, T. A. (1995). *Science*, **268**, 1721–1727.
Garman, E. (1999). *Acta Cryst.* D**55**, 1641–1653.
González, A., Pédelacq, J.-D., Sola, M., Gomis-Rüth, F. X., Coll, M., Samama, J.-P. & Benini, S. (1999). *Acta Cryst.* D**55**, 1449–1458.
Gu, Y. X., Liu, Y. D., Hao, Q., Ealick, S. E. & Fan, H. F. (2001). *Acta Cryst.* D**57**, 250–253.
Hädener, A., Matzinger, P. K., Battersby, A. R., McSweeney, S., Thompson, A. W., Hammersley, A. P., Harrop, S. J., Cassetta, A., Deacon, A., Hunter, W. N., Nieh, Y. P., Raftery, J., Hunter, N. & Helliwell, J. R. (1999). *Acta Cryst.* D**55**, 631–643.
Hauptman, H. (1997). *Curr. Opin. Struct. Biol.* **7**, 672–680.
Henderson, R. (1990). *Proc. R. Soc. London Ser. B*, **241**, 6–8.
Hendrickson, W. A. (1999). *J. Synchrotron Rad.* **6**, 845–851.
Hendrickson, W. A. & Ogata, C. M. (1997). *Methods Enzymol.* **276**, 326–344.
Hendrickson, W. A. & Teeter, M. (1981). *Nature (London)*, **290**, 107–113.
Jaskólski, M. & Wlodawer, A. (1996). *Acta Cryst.* D**52**, 1075–1081.
Kuhn, P. & Soltis, S. M. (2001). *Nucl. Instrum Methods A*, **467**, 1363–1366.
Langs, D. A., Blessing, R. H. & Guo, D. Y. (1999). *Acta Cryst.* A**55**, 755–760.
Leiros, H.-K. S., McSweeney, S. M. & Smalås, A. O. (2001). *Acta Cryst.* D**57**, 488–497.
Lesley, S. A. *et al.* (2002). *Proc. Natl Acad. Sci. USA*, **99**, 11664–11669.
Leslie, A. G. W. (1991). *Crystallographic Computing V*, edited by D. Moras, A. D. Podjarny & J. C. Thierry, pp. 27–38. Oxford University Press.
Morris, R. J., Perrakis, A. & Lamzin, V. S. (2002). *Acta Cryst.* D**58**, 968–975.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Okaya, Y. & Pepinsky, R. (1956). *Phys. Rev.* **103**, 1645–1647.
Perrakis, A., Sixma, T. K., Wilson, K. S. & Lamzin, V. S. (1997). *Acta Cryst.* D**53**, 448–455.
Peterson, M. R., Harrop, S. J., McSweeney, S. M., Leonard, G. A., Thompson, A. W., Hunter, W. N. & Helliwell, J. R. (1996). *J. Synchrotron Rad.* **3**, 24–34.
Pohl, E., González, A., Hermes, C. & van Silfhout, R. G. (2001). *J. Synchrotron Rad.* **8**, 1113–1120.
Rice, L. M., Earnest, T. N. & Brünger, A. T. (2000). *Acta Cryst.* D**56**, 1413–1420.
Roth, M., Carpentier, P., Kaikati, O., Joly, J., Charrault, P., Pirocchi, M., Kahn, R., Fanchon, E., Jacquamet, L., Borel, F., Bertoni, A., Israel-Gouy, P. & Ferrer J.-L. (2002). *Acta Cryst.* D**58**, 805–814.
Smith, J. L., Thompson, A. & Ogata, C. M. (1996). *Synchrotron Radiat. News*, **9**, 12–14.
Schmidt, A., González, A., Morris, R. J., Costabel, M., Alzari, P. M. & Lamzin, V. S. (2002). *Acta Cryst.* D**58**, 1433–1441.
Terwilliger, T. C. (2000). *Acta Cryst.* D**56**, 965–972.
Terwilliger, T. C. & Berendzen, J. (1997). *Acta Cryst.* D**53**, 571–579.
Thompson, A. (1997). *Proceedings of the CCP4 Study Weekend. Recent Advances in Phasing*, edited by K. S. Wilson, G. Davies, A. W. Ashton & S. Bailey, pp. 97–102. Warrington: Daresbury Laboratory.
Wang, B.-C. (1985). *Methods Enzymol.*, **115**, 90–112.
Yang, C. & Pflugrath, J. W. (2001). *Acta Cryst.* D**57**, 1480–1490.