Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Crystallization and preliminary crystallographic analysis of phosphonoacetaldehyde hydrolase

Marc C. Morais,[a] Angela S. Baker,[b] Debra Dunaway-Mariano[b] and Karen N. Allen[a]*

[a]Department of Physiology, Boston University School of Medicine, Boston, Massachusetts 02118-2394, USA, and [b]Department of Chemistry, University of New Mexico, Albuquerque, New Mexico 87131-1096, USA

Correspondence e-mail: allen@med-xtal.bu.edu

Received 21 July 1999
Accepted 7 December 1999

Phosphonoacetaldehyde hydrolase, a C—P bond-cleaving enzyme which utilizes an unusual bicovalent catalytic strategy, has been crystallized by the hanging-drop vapor-diffusion method using PEG 4000 as the precipitant. The crystals belong to the monoclinic system and belong to space group $C2$, with unit-cell parameters a = 210.5, b = 45.5, c = 64.7 Å, β = 105.0°. The asymmetric unit contains a dimer related by a non-crystallographic dyad. In addition to a 2.7 Å native data set, the following data sets have been collected: a 2.4 Å data set from crystals complexed with the intermediate analog vinyl sulfonate, a 3.0 Å three-wavelength MAD data set from crystals complexed with the product analog WO_4^{2-}, as well as several heavy-atom data sets to 3.0 Å or better, of which only three have proven useful for MIR calculations. Examination of the native Patterson map revealed NCS that made previously uninterpretable derivative data useful. Independent phase sets were first calculated and refined for the MAD and MIR experiments separately and were then combined. The combined phase set was further improved by solvent flattening, histogram matching and NCS averaging. Interpretation of the resulting electron-density map is currently under way.

1. Introduction

In recent years, it has become increasingly clear that phosphonates (compounds possessing a covalent carbon–phosphorous bond) are both more common and more important in nature than originally thought (Lee *et al.*, 1992). Certain bacteria (*e.g. Pseudomonas aeruginosa, Bacillus cereus* and *Salmonella typhimirium*) have evolved a two-step metabolic pathway for the degradation of 2-aminoethylphosphonate (2-AEP; La Nauze *et al.*, 1968; Lacoste & Neuzil, 1969), the most ubiquitous of the naturally occurring phosphonates. The first step involves a transamination of 2-AEP to yield phosphonoacetaldehyde. In the second step, phosphonoacetaldehyde hydrolase (phosphonatase), the subject of this paper, hydrolyzes phosphonoacetaldehyde to acetaldehyde and inorganic phosphate. Because the C—P bond is relatively resistant to hydrolysis and as higher plants and animals are unable to cleave this bond, it is believed that organisms capable of degrading phosphonates may play an important role in the phosphorus cycle in nature (Rosenberg & La Nauze, 1967). In addition to their role in nature, enzymes capable of hydrolyzing C—P bonds are interesting from an industrial/agricultural perspective. Phosphonates are commonly used as insecticides, herbicides and starting materials in various organic syntheses. Enzymes engineered to biodegrade a broad range of such compounds may provide a convenient way to detoxify industrial/agricultural phosphonates.

Sequence analysis of phosphonatase indicates that it belongs to a newly discovered family of proteins, the HAD (haloacid dehalogenase) superfamily (Koonin & Tatusov, 1994; Baker *et al.*, 1998). Other members of the hydrolase family include several phosphotransfer enzymes such as β-phosphoglucomutase, phosphoserine phosphatase and the P-type ATPases (Aravind *et al.*, 1998), as well as enzymes carrying out related chemistry such as L-2-haloacid dehalogenase. All members of the HAD family possess a conserved aspartate residue which is postulated to participate in hydrolysis of the bond in question. In phosphonatase, this conserved aspartate is believed to be the nucleophile which attacks the P atom of the phosphonoacetaldehyde Schiff-base intermediate (Baker *et al.*, 1998). If such a mechanism is correct, then phosphonatase would be the first example of an enzyme using bicovalent catalysis on a single substrate. At present, only the structures of the L-2-haloacid dehalogenase from *Pseudomonas sp*. YL and *Xanthobacter autotrophicus* (Hisano *et al.*,

Table 1
Crystallographic statistics.

Crystal	Resolution (Å)	No. of unique reflections	Percentage complete	R_{merge}†	$\langle I/\sigma \rangle$
Native	2.7	13003	85.0 (57.6)‡	11.1	8.5 (2.0)‡
Vinyl sulfonate	2.4	20810	96.7 (99.3)‡	8.7	11.5 (3.1)‡

Heavy-atom statistics for MIR phasing

Derivative	R_{iso}§	No. of sites	Resolution range (Å)	R_c¶	f_h/E†† (acentrics)	f_h/E†† (centrics)	Anomalous data?
W	0.179	2	20.0–3.0	0.72	0.61	0.68	No
Hg	0.150	2	20.0–3.0	0.72	0.52	0.54	No
Ir	0.179	2	20.0–3.0	0.56	1.23	1.36	No

Heavy-atom statistics for MAD phasing of tungstate derivative

λ_1 (Å)	1.2146398
λ_2 (Å)	1.2142318
λ_3 (Å)	1.196994
Resolution (Å)	3.0
Completeness (%)	λ_1 = 94 (91), λ_2 = 94 (91), λ_3 = 94 (91)
R for dispersive differences	$\lambda_2 - \lambda_1 = 0.049$, $\lambda_3 - \lambda_1 = 0.056$, $\lambda_3 - \lambda_2 = 0.048$
R for anomalous differences	$\lambda_1 = 0.088$, $\lambda_2 = 0.106$, $\lambda_3 = 0.079$
Overall figure of merit for MAD phasing (before density modification)	0.41 (0.59–4.0 Å)

† $R_{merge} = 100 \sum_h \sum_i (|F_{hi}^2 - \langle F_h^2 \rangle|) / \sum_h \sum_i$, where F_{hi}^2 is the square of the *i*th intensity measurement of reflection *h* and $\langle F_h^2 \rangle$ is the mean-squared intensity of the reflection. ‡ Numbers in parentheses represent the value for the highest resolution shell. § R_{iso} = mean fractional isomorphous difference = $\sum |F_{ph} - F_p| / \sum F_p$, where F_{ph} and F_p are the derivative and native structure factors, respectively. For MAD analysis, *R* is calculated between data sets at two different wavelengths for dispersive differences or between F^+ and F^- at a particular wavelength for anomalous differences. ¶ $R_c = \sum |F_{obs}^H - F_{calc}^H| / F_{obs}^H$, where F_{obs}^H and F_{calc}^H are the observed and calculated structure factors, respectively. †† f_h/E = phasing power, where f_h is the root-mean-square calculated heavy-atom structure factor and *E* is the root-mean-square lack-of-closure error.

1996; Ridder *et al.*, 1997) and the non-catalytic domain of epoxide hydrolase (Argiriadi *et al.*, 1999) are known to atomic resolution from the superfamily members. It is therefore important to determine the three-dimensional structure of other members of the HAD superfamily, such as phosphonatase, in order to further explore the ways in which related catalytic chemistry and unique substrate specificity evolves from a common protein scaffold.

2. Materials and methods

Phosphonatase from *B. cereus* expressed in *Escherichia coli* was purified as described previously (Baker *et al.*, 1998). The protein was stored in 50 m*M* HEPES, 10 m*M* $MgCl_2$, 0.1 m*M* DTT buffer at 277 K. Prior to crystallization, the protein was transferred to a 1 m*M* HEPES, 10 m*M* $MgCl_2$, 0.1 m*M* DTT solution by repeated dilution and concentration in a Millipore Ultrafree concentrator. The final protein concentration of 10 mg ml^{-1} was determined by a Bradford protein assay (BioRad). Initial crystallization trials were conducted by the sparse-matrix screening method (Jancarik & Kim, 1991) with hanging-drop vapor-diffusion geometry, using Crystal Screens I and II (Hampton Research). Crystals were first obtained from 30% PEG 4000, 100 m*M* Tris–HCl pH 8.5, 10 m*M* $MgCl_2$ using equal volumes (2.5 µl) of the described protein solution and well solution. These conditions were optimized to 30% PEG 4000, 100 m*M* Tris–HCl pH 7.4, 100 m*M* $MgCl_2$ using equal volumes (10 µl) each of protein solution and well solution. Crystals usually grow in one week, but sometimes take more than four weeks to grow. Crystal density measurements were made using the Ficoll gradient method (Westbrook, 1985). The crystal density and unit-cell parameters are consistent with a dimer in the asymmetric unit (Matthews, 1985). A dimer in the asymmetric unit is also consistent with gel-filtration studies, which indicate that phosphonatase exists as a homodimer in solution (La Nauze & Rosenberg, 1968).

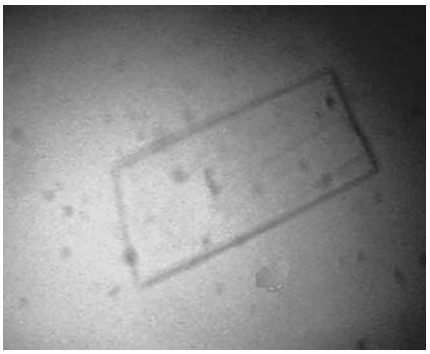

Figure 1
A crystal of phosphonoacetaldehyde hydrolase from *B. cereus* grown from a PEG 4000 solution.

Prior to data collection, glycerol was added as a cryoprotectant to the drop until a final concentration of ∼20% was reached and the crystals were allowed to soak for 1–24 h. Crystals were frozen directly for data collection in a stream of nitrogen gas cooled to 93 K with liquid nitrogen. The 2.7 Å native data set and all heavy-atom derivatives used in MIR phasing were collected on an R-AXIS II area detector using monochromatic Cu $K\alpha$ radiation generated by a Rigaku RU300 rotating-anode source. The 2.3 Å data set from phosphonatase complexed with vinyl sulfonate was collected on beamline X12B at Brookhaven National Laboratory's National Synchrotron Light Source using a 60 mm MAR detector. The three-wavelength MAD data set was collected on beamline X4A at Brookhaven National Laboratory's National Synchrotron Light Source using Fuji image plates. The three wavelengths were chosen to maximize both the anomalous scattering at each wavelength and the dispersive difference between wavelengths.

The *DENZO* and *SCALEPACK* packages (Otwinowski, 1990) were used for determination of unit-cell parameters, data indexing, reduction and scaling. Patterson maps were initially generated and solved using the *XTALVIEW* package (McRee, 1993). The automated program *SOLVE* (Terwilliger & Berendzen, 1999) was used to confirm and refine heavy-atom sites as well as to find additional sites. Density modification, map calculation and other data manipulations were carried out using the *CCP*4 program suite (Collaborative Computational Project, Number 4, 1994). The graphics program *O* (Jones *et al.*, 1991) was used for map examination.

3. Results and discussion

A systematic search for crystallization conditions produced plate-like crystals which diffracted to better than 2.7 Å and were suitable for further X-ray analysis (Fig. 1). The optimized crystallization conditions were 30% PEG 4000, 100 m*M* Tris–HCl pH 7.4, 10 m*M* $MgCl_2$. A native data set to 2.7 Å resolution was collected from such crystals. The data were indexed, reduced and scaled automatically using the programs *DENZO* and *SCALEPACK*. The crystals are monoclinic, space group *C*2, with unit-cell parameters a = 210.5, b = 45.5, c = 64.7 Å, β = 105.0°. Inspection of the diffraction pattern using the *CCP*4 program *HKLVIEW* showed that reflections with

$h + k = 2n + 1$ are absent, confirming that the space group is $C2$. Crystal density measurements using the Ficoll gradient method (Westbrook, 1985) give a solvent content of 55% (Matthews, 1985). The crystal density and unit-cell parameters are consistent with a dimer in the asymmetric unit. Addition of 5 mM of the intermediate analog vinyl sulfonate to the crystallization conditions described above resulted in larger thicker crystals which diffract to better than 2.1 Å resolution. A complete data set to 2.3 Å was collected from crystals co-crystallized with vinyl sulfonate. Comparable unit-cell parameters suggest that these crystals are isomorphous with native crystals. Table 1 lists relevant statistics for the two data sets.

Despite the fact that phosphonatase shows limited sequence identity (10–15%) to 2-L-haloacid dehalogenase, a member of the HAD superfamily whose structure is known to atomic resolution (Hisano *et al.*, 1996), attempts to use this protein as a search model in molecular replacement met with little success. It was therefore necessary to begin screening for heavy-atom derivatives. Data sets to at least 3 Å were collected on crystals soaked in the following solutions: 10 mM $Na_2WO_4^{2-}$, 1 mM ethylmercury phosphate (EMP), 10 mM $TbCl_3$, 1 mM uranyl acetate, 1 mM potassium hexachloroiridate and 1 mM trimethyllead acetate. One complication that arose during phase determination using heavy-atom data sets is the fact that the unit-cell length c doubles upon derivatization, giving unit-cell parameters a = 210, b = 45, c = 129 Å, $\beta = 104.96^\circ$ and a tetramer in the asymmetric unit. The simplest interpretation of this change is that heavy-atom binding changes the relative orientation of two dimers. Thus, what was once a crystallographic translation along the c axis of one unit-cell length in the smaller unit cell is now a non-crystallographic pseudo-translation along c of half a unit-cell length of the larger unit cell. A pseudo-translation of half a cell length can be thought of as arising from a non-crystallographic dyad running parallel to a crystallographic twofold axis (Epp *et al.*, 1971). An NCS dyad of this type manifests itself in reciprocal space as systematic low intensity of every other reflection along the direction corresponding to the direction of pseudo-translation in real space (l in our case; Epp *et al.*, 1971). Examination of heavy-atom data revealed that all odd-numbered reflections along l were indeed either very weak or completely absent. The location of the NCS twofold axis parallel to the crystallographic twofold can be found by examining a native Patterson map of the data set in question (Epp *et al.*, 1971). Examination of heavy-atom native Patterson maps (non-difference, F_{obs} used as coefficient) revealed a very large peak on the $v = 0$ Harker plane at $u = 0$, $w = \frac{1}{2}$ (Fig. 2). This corresponds to an NCS dyad parallel to b, the crystallographic twofold, which goes through the point $x = 0$, $z = \frac{1}{4}$. Rotation around such a twofold would indeed generate a pseudo-translation of $\frac{1}{2}$ along z.

The problem, in terms of using such a derivative for MIR phasing, lies in the fact that differences between the heavy-atom data set and the native data set are no longer solely a consequence of the additional scattering introduced by the heavy atom, but now arise from non-isomorphism. To circumvent this problem, one of the heavy-atom data sets indexing in the larger unit cell was chosen to function as a pseudo-native data set. To choose which heavy-atom data set to use as a pseudo-native data set, all the derivatives were first reindexed in the smaller unit cell. Next, the mean isomorphous difference between each derivative and the native was calculated. Since the trimethyllead acetate derivative had the lowest mean isomorphous difference to the native, it was selected for use as a pseudo-native data set. Once a pseudo-native data set was chosen, all subsequent calculations using any data set were performed using data indexed in the larger unit cell. Difference Patterson maps calculated between the pseudo-native data set and the WO_4^{2-} data set were interpretable and two WO_4^{2-} sites were located. Difference Patterson maps calculated for the EMP data set were also interpretable and two mercury sites were found. Difference Patterson maps calculated using the $TbCl_3$ data set and the potassium hexachloroiridate data set could not be solved by hand. However, the automated program *SOLVE* found two iridium sites that proved useful in phasing. Relevant heavy-atom statistics are listed in Table 1.

In order to further circumvent non-isomorphism problems, a three-wavelength MAD data set was collected on a crystal soaked in 10 mM WO_4^{2-}. The three wavelengths were chosen to maximize both the anomalous signal within a wavelength and the dispersive differences between wavelengths (Table 1). The crystals initially diffracted to 2.1 Å, but decayed in the beam despite being cooled to 93 K, so that a complete data set could only be collected to 3 Å. Examination of anomalous difference Patterson maps from this data set indicated that the same sites found for the WO_4^{2-} data set collected on the rotating-anode generator were present. The automated program *SOLVE* was used to confirm and refine these sites.

Initially, no special precautions were taken to account for the fact that the trimethyllead acetate was present in the pseudo-native data set but not in any of the other derivative data sets. However, the anomalous and dispersive difference

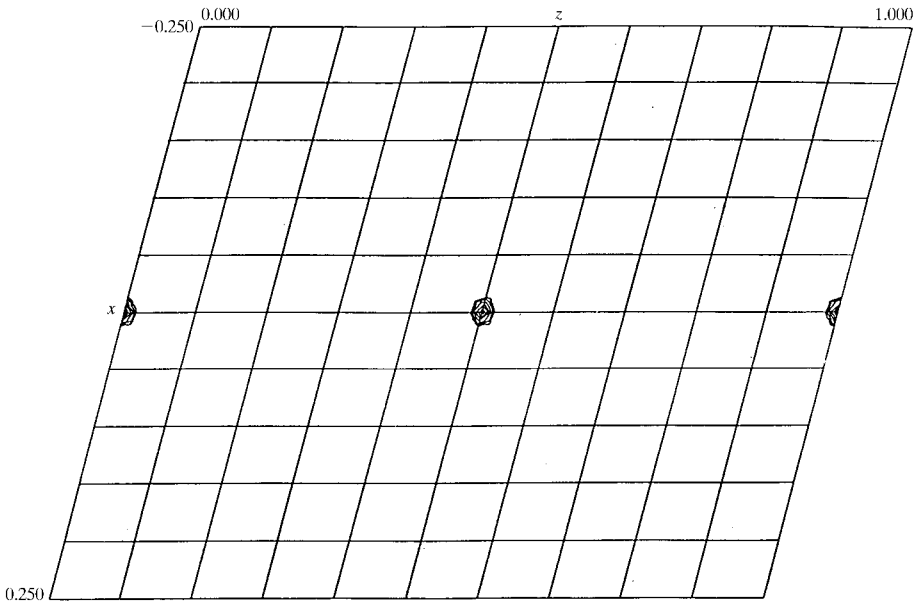

Figure 2
The $v = 0$ section of the 20.0–3.0 Å native Patterson map (non-difference, F_{obs} used as coefficients) of MAD λ_3. The map is contoured so that only the origin (at $u = 0$, $w = 0$ and $u = 0$, $w = 1$) and the pseudo-origin (at $u = 0$, $w = \frac{1}{2}$) are visible. The pseudo-origin at $u = 0$, $w = \frac{1}{2}$ corresponds to a non-crystallographic dyad parallel to y at $x = 0$, $z = \frac{1}{4}$. The intensity of the pseudo-origin peak is ∼0.5 of the origin.

Patterson maps calculated for the MAD data set yielded the same two tungstate sites as the difference Patterson using the pseudo-native data set and the tungstate data set collected in-house. This indicates that the peaks used to solve the difference Patterson for the pseudo-native data set minus the tungstate data set resulted from the presence of tungstate rather than the absence of trimethyllead acetate in the derivative data set. Additionally, we attempted to use information regarding the position of trimethyllead acetate in the pseudo-native data set to improve the MIR phase distribution. A difference Fourier map using MAD phases was calculated between λ_3 of the MAD data set and the trimethyllead acetate pseudo-native data set. A single site was found. This site did not coincide with the EMP or IR site used in the MIR experiment. To account for the absence of this site in the MIR derivatives, this site was added to each of the derivatives, but the occupancy was set to a negative value to model the absence of the lead site. This approach did not result in any noticeable improvement in map quality (*i.e.* connectivity was not improved, nor was recognition of secondary structure or protein–solvent boundaries facilitated) compared with maps calculated from the original MIR phase distribution. Thus, we believe that the effect of the lead site on the phase information is not significant.

Two approaches were taken to incorporate the information regarding the pseudo translation of $\frac{1}{2}$ on c. The first approach was to treat it as NCS, either as a pseudo-translation along c or, equivalently, as a rotation around an NCS twofold axis parallel to the crystallographic twofold axis and passing through the point $a = 0, c = \frac{1}{4}$. The alternative approach was to treat the pseudo-translation as if it were crystallographic symmetry by adding two additional origins, one of $(0, 0, \frac{1}{2})$ and the other of $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$, to the $C2$ space-group symmetry files in the *CCP*4 package and then run the *CCP*4 program *CAD*. Adding the additional origins is equivalent to treating odd-numbered l reflections as systematic absences and thus these reflections were removed by the *CAD* program. The second approach is appropriate since the size of the $u = 0, w = \frac{1}{2}$ peak on the $v = 0$ Harker plane of the native Patterson of MAD λ_3 compared with the origin suggests that this NCS symmetry is nearly crystallographic and the average signal-to-noise ratio for odd-numbered l reflections was only slightly greater than 1. The consequences of including an additional origin are that two identical dimers appear in the unit cell, rather than two distinct dimers that may differ either in their individual structures, their relative orientations in the unit cell, or both. The maps made using an additional origin were far superior in clarity and connectivity to those using NCS averaging. Thus, when this approach was used in refining heavy-atom positions and calculating initial phases and maps in the *SOLVE* package, both MIR phases and MAD phases yielded maps where the solvent envelope was readily discernible and regions of β-sheet were apparent. The MAD map alone was considerably better than the MIR map alone, since it showed additional α-helical regions.

The probability distributions from the two phase sets, MAD and MIR, were combined using the *SIGMAA* program from the *CCP*4 package. A map calculated from this combined phase set was slightly better in terms of connectivity than a map calculated from either phase set alone. Therefore, this combined phase set was subject to density modification using the program *DM* from the *CCP*4 package. It was possible to determine the location of the remaining non-crystallographic twofold in the resulting map. Additionally, a significant region of β-sheet was clearly visible, as well as several α-helices.

Further interpretation of the map is currently under way. Once the model is complete, it will be used to phase both the native data set and the vinyl sulfonate data set. Phases for these two data sets will either be calculated directly from the model or *via* the molecular-replacement method if necessary.

This work was funded by a grant from the National Science Foundation MCB 9630430 (to KNA) and the National Institutes of Health GM 28688 (to DDM). MCM was supported in part by a training grant from the National Institutes of Health GM 08291. Research carried out (in part) at beamlines X4A and X12B at the National Synchrotron Light Source, Brookhaven National Laboratory, which is supported by the US Department of Energy, Division of Materials Sciences and Division of Chemical Sciences under contract number DE-AC02-98CH10886. We thank Craig Ogata for helpful advice on MAD data collection.

References

Aravind, L., Galperin, M. Y. & Koonin, E. V. (1998). *Trends Biochem. Sci.* **23**, 127–129.
Argiriadi, M. A., Morisseau, C., Hammock, B. D. & Christianson, D. W. (1999). *Proc. Natl Acad. Sci. USA*, **96**, 10637–10642.
Baker, A. S., Ciocci, M. J., Metcalf, W. W., Kim, J., Babbitt, P. C., Wanner, B. L., Martin, B. M. & Dunaway-Mariano, D. (1998). *Biochemistry*, **37**(26), 9305–9315.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Epp, O., Steigeman, W., Formanek, H. & Huber, R. (1971). *Eur. J. Biochem.* **20**, 432–437.
Hisano, T., Hata, Y., Fujii, T., Liu, J.-Q., Kurihara, T., Esaki, N. & Soda, K. (1996). *J. Biol. Chem.* **271**, 20322–20330.
Jancarik, J. & Kim, S.-H. (1991). *J. Appl. Cryst.* **24**, 409–411.
Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Koonin, E. V. & Tatusov, R. L. (1994). *J. Mol. Biol.* **244**, 125–132.
Lacoste, A. M. & Neuzil, E. (1969). *C. R. Acad. Sci. Ser. D*, **269**, 254–257.
La Nauze, J. M. & Rosenberg, H. (1968). *Biochim. Biophys. Acta*, **165**, 438–447.
Lee, S.-L., Hepburn, T. W., Swartz, W. H., Ammon, H. L., Mariano, P. S. & Dunaway-Mariano, D. (1992). *J. Am. Chem. Soc.* **114**, 7346.
McRee, D. E. (1993). *Practical Protein Crystallography*. Boston: Academic Press.
Matthews, B. W. (1985). *Methods Enzymol.* **114**, 176–187.
Otwinowski, Z. (1990). *DENZO Data Processing Package*. Yale University, New Haven, Connecticut, USA.
Ridder, I. S., Rozeboom, H. J., Kalk, K. H., Janssen, D. B. & Dijkstra, B. W. (1997). *J. Biol. Chem.* **272**, 33015–33022.
Rosenberg, H. & La Nauze, J. M. (1967). *Biochim. Biophys. Acta*, **141**, 79–90.
Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 849–861.
Westbrook, E. M. (1985). *Methods Enzymol.* **114**, 187–196.