Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Solution of the structure of the cofactor-binding fragment of CysB: a struggle against non-isomorphism

Koen H. G. Verschueren,[a]* Richard Tyrrell,[a]† Garib N. Murshudov,[a,b] Eleanor J. Dodson[a] and Anthony J. Wilkinson[a]

[a]Protein Structure Research Group, Department of Chemistry, University of York, Heslington, York YO1 5DD, England, and [b]CLRC Daresbury Laboratory, Warrington WA4 4AD, England

† Present address: NIMR, The Ridgeway, London NW7 1AA, England

Correspondence e-mail: koen@yorvic.york.ac.uk

Received 4 February 1998
Accepted 10 July 1998

The elucidation of the structure of CysB(88–324) by multiple isomorphous replacement (MIR) techniques was seriously delayed by problems encountered at every stage of the analysis. There was extensive non-isomorphism both between different native crystals and between native and heavy-atom-soaked crystals. The heavy-atom substitution was invariably weak and different soaking experiments frequently led to substitution at common sites. These correlated heavy-atom binding sites resulted in an overestimation of the phase information. Missing low-resolution reflections in the native data set, constituting only 2% of the total observations, reduced the power of density modification and phase refinement. Finally, the extensive dimer interface made it difficult to isolate a single molecule in the course of model building into the MIR maps. The power of maximum-likelihood refinement (*REFMAC*) was exploited in solving the structure by means of iterative cycles of refinement of a partial model, initially comprising only 30% of the protein atoms in the final coordinate set. This technique, which uses experimental phases, can automatically discriminate the correct and incorrect parts of electron-density maps and give properly weighted combined phases which are better than the experimental or calculated ones. This allowed the model to be gradually extended by manual building into improved electron-density maps. A model generated in this way, containing just 50% of the protein atoms, proved good enough to find the transformations needed for multi-crystal averaging between different crystal forms. The averaging regime improved the phasing dramatically such that the complete model could be built. The problems, final solutions and some possible causes for the observed lack of isomorphism are discussed.

1. Introduction

Despite the great advances in techniques for data collection and in crystallographic methodology made during the last decade, there are still protein crystal structures which prove extremely elusive. CysB was such a case; in order to extract some general lessons, the experiment was re-examined in the light of the known structure. The non-isomorphism is examined in detail; the reliability of phase-probability distribution in density-modification procedures is discussed; the contribution of maximum-likelihood refinement of a very partial model is evaluated as a tool for phase improvement; and the power of averaging techniques is described.

CysB is a tetramer of identical polypeptides of 324 amino-acid residues ($M_r = 36$ kDa) which acts in the regulation of the expression of the genes involved in the biosynthesis of cysteine in Gram-negative bacteria. It binds upstream of a number of *cys* promoters where it acts as a transcriptional

Table 1
Data-collection and phasing statistics of the many native data sets collected in the three different crystal forms together with those for four form *A* derivative data sets.

$R_{\rm merge} = \sum_{hkl}\sum_i |I - \langle I\rangle| / \sum_{hkl}\sum_i I$. $R_{\rm iso} = \sum_{hkl} |F_{\rm PH} - F_P| / \sum_{hkl} |F_P|$ (calculated against crystal form *A* native 1). $R_{\rm cullis} = \sum_{hkl} \big||F_{\rm PH} \pm F_{\rm P}| - F_{\rm H}({\rm calc})\big| / \sum_{hkl} |F_{\rm PH} \pm F_{\rm P}|$. Phasing power = $F(H)/E$, the mean value of the heavy-atom structure-factor amplitude divided by the residual lack-of-closure error. Completeness, overall completeness (completeness in the highest resolution shell; completeness between 20 and 8 Å). $R_{\rm merge}$ = $R_{\rm merge}$ in the highest resolution shell.

		Cell dimensions (Å)								Phasing
	Resolution				Unique	Completeness	$R_{\rm merge}$	$R_{\rm iso}$	$R_{\rm cullis}$ (centric/	power (centric/
Data set	(Å)	*a*	*b*	*c*	reflections	(%)	(%)	(%)	acentric)	acentric)
$P2_12_12$ crystal form *A*										
Native 1	10–2.12	66.58	107.97	32.83	13 799	98.2 (92.4) (45.3)†	8.1 (32.6)			
Native 2	20–1.80	66.38	107.53	32.74	20 911	93.0 (84.9) (70.6)	4.4 (31.9)	9.3		
Native 3	20–2.15	67.60	109.28	33.14	13 500	96.6 (90.6) (97.0)	9.3 (44.9)	6.7		
Composite native 1	20–2.12	66.58	107.97	32.83	13 992	99.5				
Composite native 2	20–1.80	66.38	107.53	32.74	21 297	98.7				
Derivatives used for phasing										
EMTS	20–3.65	66.70	109.20	33.30	2703	97.3 (88.6)	5.1 (30.2)	21.4	0.66/0.67	1.0/1.4
K_2PtCl_4 (Pt1)	30–2.76	66.54	109.06	33.11	6320	99.2 (96.7)	4.6 (28.6)	29.7	0.87/0.91	0.5/0.7
K_2PtCl_4 (Pt2)	15–4.50	67.16	110.55	33.37	1322	78.8 (69.5)	5.3 (31.7)	34.1	0.88/0.91	0.5/0.7
PbAc	20–2.18	66.32	108.82	33.13	10 844	84.7 (78.6)	7.6 (33.9)	25.6	0.77/0.81	0.7/1.1
$P2_12_12$ crystal form *B*										
Native	15–3.00	64.39	111.45	33.42	4953	96.2 (91.0)	7.9 (27.4)	44.1		
$P2_12_12_1$ crystal form *C*										
Native 1	26–3.00	76.77	108.92	60.69	10 792	92.5 (78.7)	7.6 (29.8)			
Native 2	30–3.50	75.29	107.90	60.14	5694	85.9 (86.7)	13.1 (39.0)			

† No reflections were measured between 20 and 10 Å.

activator. In contrast, it behaves as a repressor at its own promoter (reviewed by Kredich, 1996). CysB belongs to the LysR-type transcriptional regulators (LTTRs), a large family of coinciducer-responsive transcriptional activators which regulate genes of highly diverse function (Schell, 1993). The crystal structure of a large C-terminal fragment (residues 88–324) of CysB from *Klebsiella aerogenes*, containing the cofactor-binding domain but missing the DNA-binding determinants, has now been elucidated and refined against X-ray data extending to a resolution of 1.80 Å (Tyrrell *et al.*, 1997), providing the first three-dimensional structural insight into a protein belonging to the LysR family. The cofactor-binding core consists of two α/β domains connected by a two-stranded hinge region. The two domains form a solvent-inaccessible cofactor-binding pocket. In the CysB(88–324) structure, this cleft is occupied by a sulfate anion. Unlike full-length CysB, which is active as a tetramer, in the crystal structure the fragment forms a tight dimer generated by a twofold rotational axis of symmetry. Upon dimer formation $\sim$2400 Å^2 or 20% of the accessible surface of the monomer is buried. A more detailed description of the crystal structure and its biological implications is given elsewhere (Tyrrell *et al.*, 1997).

2. Materials and methods

2.1. Crystallization and data collection

The C-terminal fragment CysB(88–324) was prepared by limited chymotryptic digestion of *K. aerogenes* CysB (Tyrrell *et al.*, 1997) and good-quality crystals grew readily by vapour diffusion in hanging drops from a solution containing 6 mg ml^{-1} protein, 14% monomethylether polyethylene glycol (PEG) 750 and 0.1 *M* 2-(*N*-morpholino)ethanesulfonic acid (MES) buffer at pH 6.5 (Tyrrell *et al.*, 1994). The crystals belong to space group $P2_12_12$ and contain one monomer per asymmetric unit. They could be classified into two principal forms with unit-cell dimensions either close to a = 66.4, b = 107.5 and c = 32.7 Å (form *A*) or close to a = 64.4, b = 111.5 and c = 33.5 Å (form *B*). Although the cell dimensions of the two crystal forms do not differ very much, the crystals are not isomorphous (Table 1). Form *B* crystals were usually of poorer quality.

From single crystals in form *A*, two good native data sets were collected at room temperature on beamline X11 using synchrotron radiation (λ = 0.91 Å) and a MAR Research image-plate detector (300 mm) at the EMBL outstation at DESY Hamburg, Germany. The first was 98.0% complete in the resolution range 10–2.12 Å, but had 147 missing reflections in the 20–10 Å shell (native 1, Table 1). The second was 92.9% complete to a maximum resolution of 1.80 Å (native 2, Table 1). However, the latter was only 70.0% complete in the low-resolution range, with 84 reflections missing between 20 and 8 Å. During the initial model building, a third 96.6% complete data set extending to 2.15 Å spacing was collected from a native form-*A* crystal on the Rigaku R-axis IIC image plate at York (λ = 1.54 Å) (native 3, Table 1). Low-resolution reflections from this set were used to complete the 2.12 Å (193 added reflections) and 1.80 Å (386 added reflections) native

Table 2
Fractional coordinates of the different heavy-atom sites in crystal form *A* and their protein ligands in the CysB(88–324) structure.

		x	*y*	*z*		Occupancy	*B* factor (Å^2)
Hg	Site 1	0.242	0.108	0.876	(Cys163)	0.6	16.6
	Site 2	0.223	0.086	0.767	(Cys163, alternate conformation)	0.4	18.2
Pt1	Site 1	0.494	0.173	0.032	(His125)	0.6	9.9
Pt2	Site 1	0.533	0.219	0.916	(His125, alternate conformation)	0.7	25.6
	Site 2	0.281	0.026	0.356	(Asp249)	0.4	Not refined
Pb	Site 1	0.241	0.106	0.875	(Cys163)	0.6	12.8
	Site 2	0.221	0.085	0.766	(Cys163, alternate conformation)	0.4	18.8

Table 3
Comparison of the contributions of the different steps during the model building with the incomplete and completed native form *A* data sets..

Phase shifts and map correlations are compared with the final calculated CysB(88–324) model phases and electron-density map.

	Incomplete 2.12 Å native data set (10–2.12 Å)			Completed 2.12 Å native data set (20–2.12 Å)		
	Figure of merit	Phase error (°)	Map correlation	Figure of merit	Phase error (°)	Map correlation
MLPHARE	0.44	68.26	0.45	0.42	67.21	0.50
DM	0.67	63.67	0.56	0.73	60.54	0.66
Skeletonization	0.77	63.04	0.55	0.81	60.09	0.66
	Incomplete 1.80 Å native data set (20–1.80 Å)			Completed 1.80 Å native data set (20–1.80 Å)		
	Figure of merit	Phase error (°)	Map correlation	Figure of merit	Phase error (°)	Map correlation
MLPHARE	0.40	69.37	0.43	0.40	69.20	0.48
DM	0.53	68.08	0.55	0.59	67.09	0.61
Skeletonization	0.68	67.50	0.54	0.73	65.95	0.61
Multi-crystal averaging				0.46	58.08	0.73

data sets. This resulted in a 99.5% complete composite 2.12 Å native data set and a 98.7% complete composite 1.80 Å native data set, both of which were nearly complete in the low-resolution range (20–8 Å). These 'completed' data sets were eventually used in the model building and refinement. Details of the data-collection statistics and the agreement between data sets are presented in Table 1.

Two crystals grown from an 'aged' batch of protein had the same apparent morphology as the $P2_12_12$ CysB(88–324) crystals but turned out to be in the space group $P2_12_12_1$ with a dimer in the asymmetric unit and cell dimensions of *a* = 76.8, *b* = 108.9 and *c* = 60.7 Å (form *C*). The conditions for growth of these crystals could not be reproduced. Two native data sets extending to 3.0 and 3.5 Å spacing, respectively, were collected from these form-*C* crystals on beamline 9.6 (λ = 0.90 Å) using a MAR Research image-plate detector (300 mm) at the Daresbury SRS. The data-collection statistics are listed in Table 1.

All data sets, native as well as derivative, were processed with *DENZO* (Otwinowski & Minor, 1997) and scaled using the *CCP*4 software package (Collaborative Computational Project, Number 4, 1994) or *SCALEPACK* (Otwinowski & Minor, 1997).

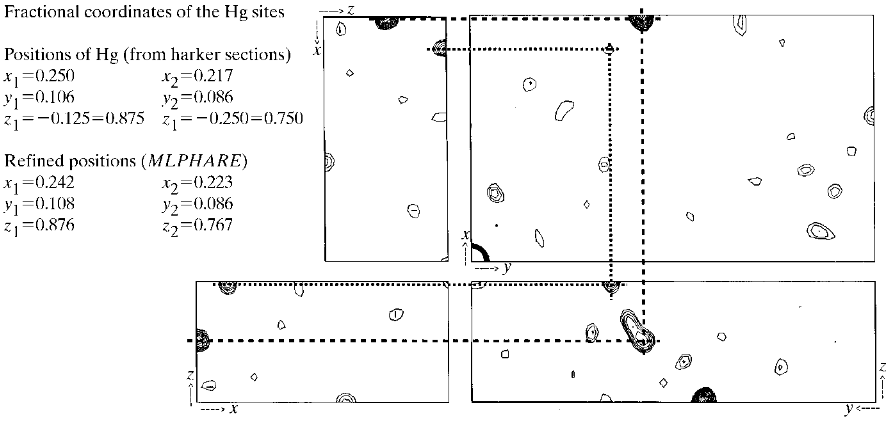

Figure 1
Harker sections at 4 Å resolution of the anomalous Patterson maps derived from the EMTS heavy-atom soak in crystal form *A*. The sections range from 0 to 0.5 of the fractional unit cell and the peaks are contoured at 0.5σ intervals starting at 2σ. The directions of the axes are indicated at the origin of the Harker sections. One of the sections is duplicated such that in space group $P2_12_12$ the cross-peaks of crystallographically related heavy-atom positions can be connected as indicated by the dashed lines.

2.2. Heavy-atom-derivative search

In total, four useful form-*A* heavy-atom-derivative data sets were collected, one after soaking in ethyl mercury thiosalicylate (EMTS), two after soaking in potassium tetrachloroplatinate(II) and one after a soak with lead(II) acetate (Table 1). Most heavy-atom-derivative data sets were collected in the laboratory on either a Rigaku R-axis IIC or a MAR Research image plate with a conventional rotating copper anode X-ray source (λ = 1.54 Å). In all cases the level of substitution was low.

2.3. Phasing and phase improvement

The EMTS derivative had two sites of substitution and gave anomalous data of sufficient quality to generate the Patterson map shown in Fig. 1. Phases calculated from these two Hg atoms were used to locate platinum and lead sites using cross-phased difference Fouriers (Table 2). The parameters of the heavy-atom sites were refined with the program *MLPHARE* (Otwinowski, 1991). These sites were cross-checked and eventually confirmed at the final stage using calculated model phases.

The MIR (multiple isomorphous replacement) phases were extended and refined to the resolution limits of the data by solvent flattening, histogram matching and Sayre's equation implemented in the program *DM* followed by 20 cycles of skeletonization (Cowtan, 1994) (Table 3).

At this stage, the additional, predominantly low-resolution reflections (§2.1) were added to the native data sets. This resulted in a considerable improvement in the final solvent-flattened maps. Although there were only small differences between the two electron-density maps, the 2.12 Å map (composite native 1, Table 1) displayed somewhat better connectivity overall and was therefore chosen for further model building.

2.4. Model building starting with concurrent refinement of partial models

The two molecules of the crystallographic dimer could not be distinguished in the electron-density map and it was impossible to trace a complete C_α chain. A partial model was built into some of the clearer regions of the map using *XAUTOFIT* (implemented in *QUANTA*; Oldfield, 1996). This partial model was then refined using *REFMAC* (Murshudov *et al.*, 1997). The maximum-likelihood residual includes a contribution from the experimental phases and the uninterpreted density was back-transformed to give an $F_{\rm part}$ to be added to the calculated structure factor (Pannu *et al.*, 1998). Such refinement is a sensitive probe of which parts of the model are correct. *REFMAC* outputs map coefficients with phases combining the experimental and calculated contributions. If the structure is correct, the electron density of the modelled sequence improves and new electron-density features become more prominent at its fringes. If the fitted sequence is incorrect, the atoms of the proposed model do not properly match the electron density calculated in the following cycle and are removed prior to rebuilding. This iterative procedure was followed until the model was able to find the correct rotation and translation matrices for transposing the model to the form-*B* and form-*C* crystals so that multi-crystal averaging (*DMMULTI*; Cowtan, 1994) could be exploited. The averaging regime improved the phases dramatically and the complete CysB(88–324) model could be built into the newly calculated electron-density maps.

3. Results

3.1. Non-isomorphism in the crystals

The lack of isomorphism between the two $P2_12_12$ crystal forms (form *A* and form *B*) is caused by a small, but significant rotation of the CysB(88–324) dimer by 3–5° around the twofold crystallographic symmetry axis (Table 4). Table 4 presents the outcome of some of the many unsuccessful soaking experiments which ended up in crystal form *B*. It also clearly shows, as expected, a correlation between the extent of rotation and the degree of non-isomorphism. Much effort was put into trying to stabilize the crystals in one particular crystal form, without success.

3.2. Screening for heavy atoms, uninterpretable Pattersons

The crystal variability within, and the non-isomorphism between, the $P2_12_12$ crystal forms made the screening for useful heavy-atom derivatives elaborate and time consuming. Only a few soaking experiments gave heavy-atom substitution in form *A*. Most of these experiments with a range of different compounds triggered the irreversible transition of crystal form *A* to the non-isomorphous form *B*.

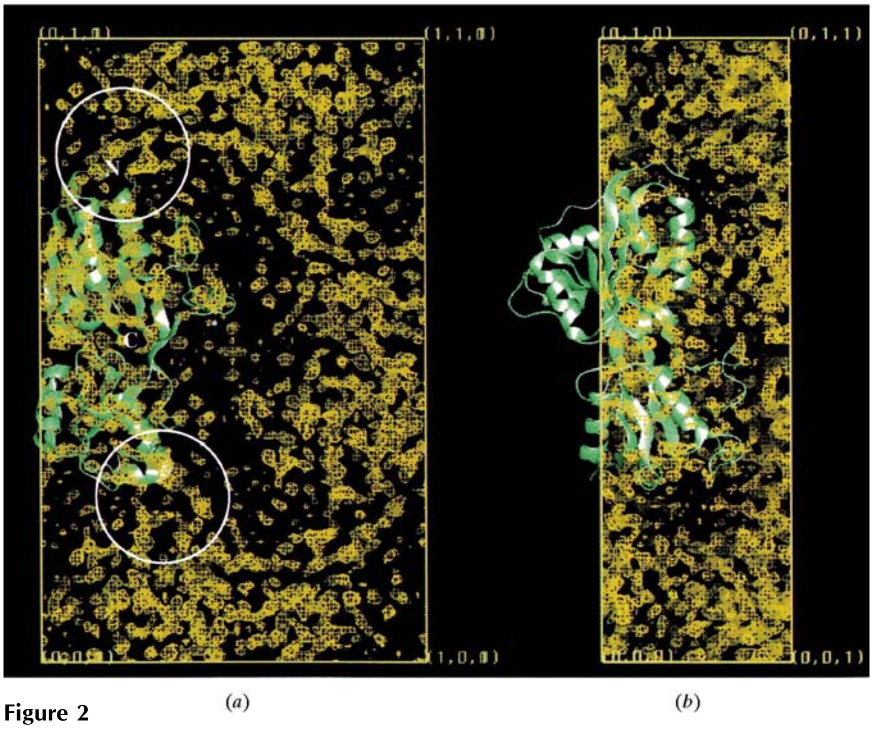

Figure 2
MIR electron-density map in the unit cell of form-*A* crystals, contoured at 1.5σ. (*a*) The view along the *z* direction, (*b*) the view along the *x* direction. A ribbon diagram of the final model of CysB(88–324) is placed in the unit cell. The N- and C-termini are labelled in (*a*). The open circles indicate the poor crystal contacts which could affect the rotation of the dimer along the twofold axis in the unit cell.

Table 4
Derivatives in the $P2_12_12$ crystal form *A* and form *B*: correlation between the rotation of the CysB(88–324) dimer after rigid-body refinement and their non-isomorphism.

$R_{\rm iso} = \sum_{hkl} |F_{\rm PH} - F_{\rm P}| / \sum_{hkl} |F_{\rm P}|$. Rotation of dimer, calculated with the crystallographic dimer in crystal form *A* native 2 as model.

	Resolution	Cell dimensions (Å)			$R_{\rm iso}$ to native 1	$R_{\rm iso}$ native	Rotation of	
	(Å)	*a*	*b*	*c*	(%)	form *B*	dimer (°)	Substitution
(*a*) Derivatives in $P2_12_12$ form *A*								
Native 1	10–2.12	66.58	107.97	32.83			−0.33	
Native 2	20–1.80	66.38	107.53	32.74	9.3			
Native 3	20–2.15	67.60	109.28	33.14	6.7		−0.04	
Used for phasing								
EMTS (Hg)†	20–3.65	66.70	109.20	33.30	21.4		0.77	Two sites Cys163
Potassium tetrachloroplatinate (Pt1)	30–2.76	66.38	109.06	33.11	29.7		1.40	One site, His125
Potassium tetrachloroplatinate (Pt2)	15–4.50	67.16	110.55	33.37	34.1		1.66	Two sites, His125, Asp249
Lead(II) acetate (Pb)	20–2.18	66.32	108.82	33.13	25.6		−1.04	Two sites, Cys163
Not used for phasing								
Erbium acetate (Er)	20–3.00	67.16	108.94	33.10	10.8		0.49	No
Potassium hexachloroiridate (Ir)	20–2.79	66.29	108.41	32.95	12.9		0.77	No
Potassium tetraoxoosmiate (Os)	20–3.30	67.62	109.66	33.31	23.0		1.50	Very weak, His127
Potassium tetraoxoosmiate (Os)	20–2.80	67.47	109.56	33.28	23.3		1.40	Very weak, His127
Selenomethionine	20–2.82	67.89	109.81	33.36	24.7		1.41	Substitution
(*b*) Derivatives in $P2_12_12$ form *B*								
Native form *B*	20–3.00	64.39	111.45	33.42	44.1		3.31	
EMTS (Hg)†	20–2.10	63.56	110.72	33.10	44.6	17.7	3.16	No
Thallium(I) acetate (Tl)	20–2.78	64.78	110.26	33.32	46.7	18.3	3.97	No
Lead(II) acetate (Pb)	20–3.00	64.81	110.33	33.38	47.7	18.5	3.65	No
Selenomethionine	20–2.99	64.50	110.90	33.20	47.5	19.3	3.67	Yes
Potassium tetrachloroplatinate (Pt)	20–3.00	65.68	110.90	33.41	49.9	20.6	4.36	No
Lanthanum(III) chloride (La)	20–3.70	64.06	110.16	33.06	49.9	21.0	4.27	Yes
Lanthanum(III) chloride (La)	20–3.29	64.02	109.88	33.09	49.4	23.2	4.32	No
Potassium tetrachloroplatinate (Pt)	10–2.80	64.47	110.06	33.24	50.4	24.8	4.26	Yes
Potassium tetrachloropalladate (Pd)	20–3.40	63.84	111.07	33.40	52.1	25.9	3.63	Yes
Dysprosium(III) chloride (Dy)	20–3.44	65.07	110.71	33.37	51.0	29.1	5.06	Yes
Holmium(III) chloride (Ho)	20–3.10	65.08	110.97	33.31	51.8	30.6	4.98	Yes
Holmium(III) chloride (Ho)	20–2.76	63.11	109.48	33.01	52.8	31.3	4.99	Yes
Terbium(III) chloride (Tb)	20–2.60	63.56	110.72	33.10	72.2	32.8	5.02	Yes
PCMB (Hg)‡	20–3.50	63.14	110.46	33.30	53.5	48.4	3.80	Yes

† EMTS = ethyl mercury thiosalicylate. ‡ PCMB = parachloromercury benzoate.

Only one interpretable set of isomorphous and anomalous Patterson maps was calculated. The EMTS derivative data set in crystal form *A* showed up two mercury sites (Fig. 1). One of these sites was also found in the anomalous Patterson map calculated from the lead(II) acetate data set. For the other useful heavy-atom data sets in form *A*, the Pattersons were featureless and the sites of substitution were identified using cross-phased difference Fouriers (Table 2). Phases calculated from the EMTS derivative were used to find the Pt and the second Pb sites. Phases derived from each of these in turn were then used to verify the heavy-atom positions in the EMTS and other derivatives.

The form-*B* crystals, which include those resulting from most of the heavy-atom-soaking experiments together with some native crystals, gave data sets that were all of lower quality so that Patterson maps were uninterpretable. In spite of this, retrospective analysis of many of these data sets by refinement of the model in form-*B* crystals revealed clear evidence of heavy-atom substitution (Table 4). With hindsight, the failure to locate these sites in the Patterson maps is the result of slight non-isomorphism even among the form-*B* crystals resulting from small rotations of the dimer about the *z* axis (data not shown). This proves to be disastrous for Patterson functions. In intensive efforts to derive experimental phases, all possible combinations of form-*B* heavy-atom-derivative and native data sets were examined in a time-consuming and ultimately unsuccessful search for an isomorphous combination.

3.3. Phase improvement and density modification

Using the phases derived from the two mercury sites, three other platinum sites (from two different platinum-soaking experiments) and two lead sites were identified in cross-

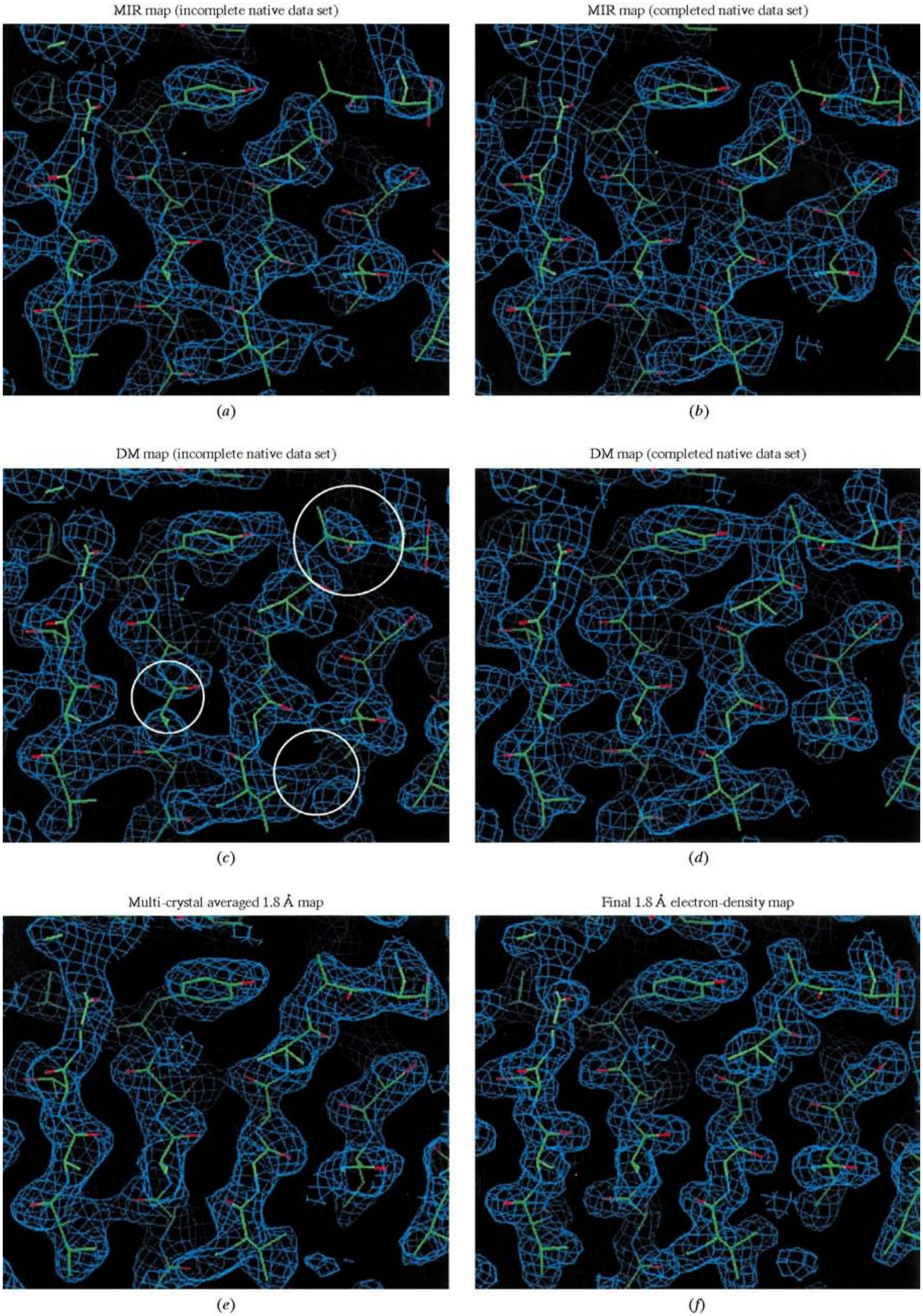
MIR map (incomplete native data set)
(a)
MIR map (completed native data set)
(b)
DM map (incomplete native data set)
(c)
DM map (completed native data set)
(d)
Multi-crystal averaged 1.8 Å map
(e)
Final 1.8 Å electron-density map
(f)

Figure 3
An illustration of the improvement of electron density after each stage in the model building and refinement. Identical regions of the electron-density maps have been displayed and are contoured at 1σ. The coordinates are from the final refined 1.8 Å model. (*a*), (*b*) MIR maps before and after completion of the native data set with the low-resolution reflections. (*c*), (*d*) *DM* maps calculated before and after completing the native data set. The white open circles highlight typical improvements in the electron-density maps brought about by completing the data set. (*e*) Electron-density map after multi-crystal averaging between crystal forms *A*, *B* and *C*. (*f*) Final 1.80 Å electron-density map.

phased difference Fourier calculations (Table 2). The two Pb atoms from the lead(II) acetate data set were located at the same positions as the Hg atoms. The F_H for Hg and Pb atoms have a correlation coefficient of 0.8, indicating that there is little additional phasing power. These lead sites were, however, included in the phase refinement since the lead(II) acetate data set extended to a much higher resolution (2.18 Å) than the other heavy-atom derivative data sets. The two positions of the Hg and Pb atoms around the Cys163 side chain (Table 2), with occupancies of 0.6 and 0.4, indicate that this residue has two alternate conformations in the crystal. Both platinum-derivative data sets also had a site in alternate positions around His125, while a third platinum was poorly substituted at Asp249. These residues are accessible from the solvent in the crystals. The combined phasing power of the derivatives was quite weak. The heavy-atom parameters, including individual B factors, were refined using *MLPHARE* (Otwinowski, 1991). The final figure of merit after phase refinement was 0.42 to a resolution of 2.18 Å (Table 3), although this was clearly an overestimation because of the correlated contribution from the Hg and Pb heavy-atom positions. The figure of merit dropped to 0.35 or 0.37, respectively, when the mercury or the lead sites were omitted from the phase calculations.

Despite the low figure of merit, a clear solvent boundary was discernible in the MIR maps calculated in the form-A unit cell (Fig. 2). It was also evident that the CysB fragment formed a closely packed dimer around the twofold axis in the crystal. The solvent-flattening, histogram-matching and Sayre's equation options implemented in *DM* (Cowtan, 1994) were used to extend and improve the MIR phases to 2.12 Å.

At this stage, completing the 2.12 Å native data set with the low-resolution reflections from an isomorphous native data set clearly made an impact on the quality of the electron-density maps as can be seen from Fig. 3. Generally interrupted density in the middle of β-strands and helices became connected, and false connections in density disappeared. The phase error was considerably reduced and the map correlations improved by ~10% upon the addition of just 193 low-resolution reflections to the 2.12 Å native 1 data set. A summary of the final phasing statistics is shown in Tables 1 and 3. The improvement in the phases, especially after density modification using the completed native data set, can be seen clearly in Fig. 4.

3.4. Model building

The first attempt at model building was carried out using the solvent-flattened electron-density maps calculated from the 2.12 Å native 1 data lacking the low-resolution reflections (20–10 Å). Although the close interaction of the molecules in the crystallographic dimer made it impossible to identify a single monomer in the maps, small sections of main chain were built into the clearer regions of electron density. The fitting procedure involved cycles of *REFMAC* to refine the partial model, interspersed with manual building and rebuilding until about 30% of the final model (mostly consisting of poly-alanine chains) had been placed into density.

After supplementing the 2.12 Å native 1 data set with missing low-resolution reflections, the newly calculated MIR and *DM* electron-density maps improved sufficiently (Figs. 3*b* and 3*d*) to allow the model to be extended using alternating cycles of model building and *REFMAC* until almost 50% of the atoms of the final model had been built.

This incomplete model proved to be sufficiently accurate to enable the successful calculation of the correct translation and rotation matrices in *AMoRe* (Navaza, 1994), essential for multi-crystal averaging (*DMMULTI*; Cowtan, 1994) between the crystal forms A, B and C. Applying these matrices to the fully refined form-A coordinate set reveals an r.m.s. deviation in main-chain-atom positions of 0.7 Å compared with the refined form-B structure, and 1.2 Å for the two molecules in the crystal form-C structure. These matrix searches were unsuccessful when using less complete search models, containing 30–45% of the atoms. However, using an 'electron-density model' created by cutting out density for a single dimer from the experimentally phased maps, *AMoRe* had, as it turned out, found a correct solution for form C, albeit with a very low correlation coefficient (11%, *cf.* background levels of 9–10%). Application of the matrix derived by this approach to the final form-A coordinate set produces a form-C model whose r.m.s. deviation from the refined form-C model is 1.3 Å. This solution was overlooked in the absence of independent criteria for assessing its quality. Had experimental phases been available in the other crystal forms, it may have been possible to refine the orientation successfully. Similarly, with a better model, rigid-body refinement may have corrected the matrices. However, without such a tool, the error in the matrices proved too great to provide any useful phase improvement.

The averaging regime improved the phases and electron density sufficiently (Figs. 3*e* and 4) for model building to be extended throughout the whole of the cofactor-binding domain of CysB. A further cycle of *REFMAC* combined with *ARP* (Lamzin & Wilson, 1993) led to the emergence of clear density for many of the side chains in the poly-alanine stretches, allowing the elucidation of the structure of CysB(88–324)

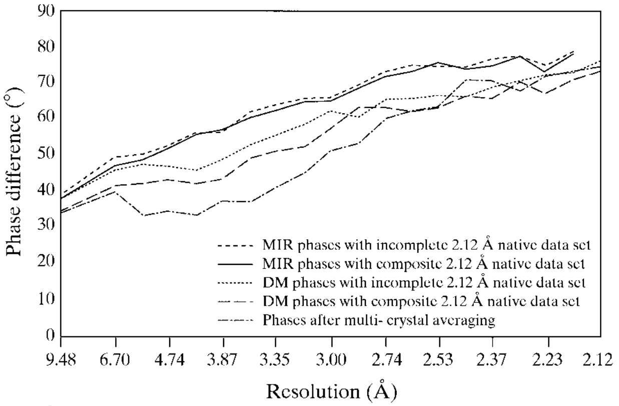

Figure 4
A comparison of the various phase sets used in the course of model building and refinement with the phases calculated from the refined CysB(88–324) model. The phase difference is plotted as a function of resolution in the range between 10 and 2.12 Å.

to be completed. The structure was subsequently refined against the original form-A 1.80 Å native 2 data set to a final R factor and free R factor of 17.9 and 24.6%, respectively. The final model contains 237 residues, 253 water molecules and one sulfate anion. The model displays good stereochemical parameters and the r.m.s. deviations from ideal values for bond lengths are 0.014 Å and for bond angles are 1.9°. A full description of the structure of CysB(88–324) and its implications for the mechanism of DNA binding and regulation of transcription is described elsewhere (Tyrrell *et al.*, 1997, and associated PDB entry 1AL3).

4. Discussion: evaluation of the procedure

With the structure of CysB(88–324) finally solved, it is possible to revisit the different steps in the structure elucidation and evaluate the contribution of each stage in the process towards the final structure solution. Analysis and refinement of many structures in crystal forms A and B give us an idea of the origin and extent of non-isomorphism between these two $P2_12_12$ crystal forms (Table 4). Calculating the accuracy of phases and the correlations of electron density after each stage with those of the final model gives us an estimate of the contribution of the different methodologies used (Fig. 4, Table 3).

4.1. Non-isomorphism between the different classes of crystals

The lack of isomorphism between the different crystals of CysB, even though sometimes very slight, was one of the major obstacles in the elucidation of the structure of the protein. CysB(88–324) forms a tight crystallographic dimer around the twofold axis in the $P2_12_12$ unit cell. The non-isomorphism between the crystal forms A and B is generated by small but significant rotations of the dimer around this axis. Originally it was thought that the binding of heavy-atom compounds induced the irreversible transition between the crystal forms from A to B. However, this may well not be the primary reason. It is possible that the form-A crystals have an inherent instability caused by weak intermolecular contacts between the N-terminal residues of one molecule and a rather flexible loop of the neighbouring molecule (highlighted in Fig. 2 by the circles), which determine the orientation of the crystallographic dimer around the twofold axis of the $P2_12_12$ form-A crystals. These contacts are broken in the form-B crystals upon rotation. This intrinsic flexibility manifests itself in small rotations, even among the different $P2_12_12$ form-A crystals (Table 4). The soaking of the crystals in the presence of heavy metals might then exacerbate this rotation of the dimer. The notion of instability inherent in the crystals is supported by two observations. Firstly, CysB(88–234) from a single batch crystallized in both crystal forms and, secondly, a few data sets (native as well as derivative) have been collected from crystals which transformed from form A to form B gradually in the course of data collection, without serious loss of crystal quality.

The extent of rotation of the dimer within the crystals limits the range of isomorphism (Table 4). Small rotations (less than 2°, crystal form A) of the dimer in the crystal still gave isomorphous data sets (which could be scaled with one another) although the phase information must be somewhat dampened by these small rotations. Another set of crystals falls in the range of dimer rotations between 3 and 5° (non-isomorphous crystal form B). Larger rotations still would presumably result in the crystals being shattered.

4.2. Pseudo-isomorphism within the form A crystals and the effect on the derivatives

The partial substitution of the heavy atoms and the small rotations in the $P2_12_12$ form-A crystals weakened the phase information from the four derivatives. Moreover, it probably obscured some of the other potential form-A derivative data sets (*e.g.* osmium, selenomethionine; Table 4). Most of the Patterson maps produced from the heavy-atom-soaked crystals that stayed in crystal form A were featureless.

Much time and effort was spent in unsuccessful attempts to improve the existing MIR phases in a search for further heavy-atom derivatives. A selenomethionine-substituted CysB protein which could be used either as a conventional derivative or in multiple-wavelength anomalous dispersion (MAD) experiments was also prepared. This protein was purified and crystallized under similar conditions to wild-type CysB(88–324) in both forms A and B, but neither direct-methods nor difference Fourier calculations using the MIR phases were successful in identifying the selenium sites, although retrospective analysis of these selenomethionine data sets indicates that full substitution at the methionine residues had taken place. Subsequent phase-refinement tests using these selenium sites indicated that the selenomethionine derivative data sets did not contribute to the overall improvement of the phases, and gave Cullis R factors of 0.93 and 0.96 and phasing powers of 0.3 and 0.4 for the centric and acentric reflections, respectively. This implies that any possible contribution of the selenomethionine substitutions to the phasing was overshadowed by the lack of isomorphism to the native crystals.

4.3. Reliability of phase-probability distributions

It is important to have reliable phase-probability distributions not only for calculating the initial map, but also for the weighting of the map coefficients used in density modification and maximum-likelihood refinement.

The particular problem associated with phasing when the only useful derivatives are strongly correlated, is overestimation of the initial phase probabilities. This in turn reduces the effectiveness of density-modification procedures, which accept the overweighted phases as correct estimates which must not be altered. To assess the seriousness of this, test calculations with *DM* were carried out. Each run was repeated using the same phases with different phase-probability distributions imposed. The solvent-flattening and histogram-matching options of *DM* were used (Cowtan & Main, 1996) with the unimodal version of the probability distribution derived from the figure of merit and centroid phase. For this particular test, phases derived from the lead(II) acetate derivative alone were used. The figure of merit is meant to be an estimate of cos $\Delta\varphi$, where $\Delta\varphi$ is the phase difference between the true phase and the calculated one. For test purposes, an 'ideal' figure of merit equal to $\cos(\varphi_{\rm MIR} - \varphi_{\rm Final\ model})$ was generated.

The initial average phase difference, for reflections in the range 20–2.18 Å resolution, between the single isomorphous replacement with anomalous scattering (SIRAS) phase set and the final one, is 69°. Running *DM* with the figure of merit derived from *MLPHARE* improves the phases marginally, reducing the average phase difference to 67° for all reflections. Replacing the SIRAS figure of merit with max(0,cos$\Delta\varphi$) resulted in a dramatic improvement. *DM* reduced the phase error to 41° (48° when solvent flattening alone was used and 44° when the histogram-matching option alone was applied) for the reflection set. Even replacing the figure of merit by max[FOM$_{\rm old}$/2,(FOM$_{\rm old}$ + cos$\Delta\varphi$)/2], where FOM = figure of merit, improved the performance, reducing the phase error to 52°.

To test the extent to which density modification is biased towards the existing weighted phases, the tests were repeated, replacing the figure of merit with max(0,cos$\Delta\varphi$) where the $\Delta\varphi$ was derived from an early set of unrefined coordinates containing only 45% of the atoms. In this case, the difference between the phases derived from this incomplete model and the phases after *DM* fell to 53°. The electron-density map calculated using these phases more closely fitted the 45% model than the final model, showing that the phase-probability distribution strongly influences the phase error after density modifications.

It is possible to postulate why density-modification procedures might be misled. First of all, the initial density and solvent boundary are distorted both by errors in phases and errors in the figures of merit. A satisfactory solvent outline can be obtained from quite a small subset of correct phases, but this becomes systematically 'blurred' as incorrect structure factors are added. Secondly, density modification is a cyclic procedure, combining phases calculated from modified electron density with the original ones, using weights based on both the original and the modified phase-probability distributions. Obviously, the associated sources of error are not independent and although algorithms exist which improve the situation substantially (Abrahams, 1997; Cowtan & Main, 1996) the problem of recombining phases is not completely solved.

In principle, one heavy-atom derivative should be sufficient for structure solution, providing that the heavy-atom refinement and phasing programs give reliable probability distributions. For this to be achieved, a proper understanding of the sources of errors and their treatment is needed. These sources of error include non-isomorphism, inaccurate heavy-atom parameters and uncertainty in wavelength for MAD/MIRAS (multiple isomorphous replacement with anomalous scattering) experiments. The maximum-likelihood approach, developed by de la Fortelle & Bricogne (1997) and implemented in *SHARP*, provides, in principle, a better approach for heavy-atom refinement and phasing which should improve the reliability of the phase-probability distribution, resulting in more effective density-modification procedures. One more problem in heavy-atom refinement programs is dealing with correlated heavy-atom derivatives. Terwilliger & Berendzen (1996) discussed an approach to solving this problem. However, a more robust approach may require the shape of the likelihood function being changed so as to take into account the correlation between different derivatives.

Similarly, 'blurring' the phase-probability distribution in maximum-likelihood refinement (Pannu *et al.*, 1998) was influential in compensating for the overestimation of experimental phase accuracy caused by the correlated heavy-atom derivatives. Even with a very partial and incorrect model, the contribution to the calculated structure factors from back-transformed uninterpreted electron density enables *REFMAC* to distinguish between the correctly and incorrectly built pieces of the model, and to generate appropriately weighted combined phases which are better than either the experimental phases or the calculated ones. This resulted in improved density in those parts of the map which did not yet contain any phase information.

4.4. Influence of the low-resolution reflections

Another important factor in the successful elucidation of the CysB(88–324) structure was the completion of the native data sets with the low-resolution reflections which considerably improved the electron-density maps, even though these reflections constitute less than 2% of the total number of those observed.

The completeness of the data, especially at low resolution, is important for density-modification procedures which rely on all reflections and the relationships among them. The strong low-resolution reflections are, therefore, of vital importance in the success of such methods. The figure of merit, phase error and map correlations all improved upon filling the low-resolution shells with these missing reflections (Table 3, Fig. 4).

4.5. Multi-crystal averaging

Multi-crystal averaging is well known to be a very powerful tool for phase improvement but it can only be used after the appropriate transformation matrix from one crystal form to the other is known. This matrix could only be determined from a sufficiently complete search model. The correct solution must stand out from the background, as slightly less complete models resulted in molecular replacement calculations showing no contrast between the correct answer and the background.

5. Conclusions

The determination of the three-dimensional structure of CysB(88–324) presented here is a case study of how lack of good phase information can be overcome. This deficiency was mainly caused by non-isomorphism between the different crystals, not only following soaking of crystals with heavy-atom derivatives, but also between the different native crystals. This, combined with the partial occupancies and common positions of the heavy metals, led to the weak phasing power of the derivatives. The structure was solved by means of an elaborate combination of MIR techniques, the building of a partial model and its gradual extension from 30 to 50% of the total structure, phase improvement using maximum-likelihood refinement techniques, and finally multi-crystal averaging among three different crystal forms. The study illustrates the pitfalls that can be encountered and examines what general lessons can be drawn.

(i) Non-isomorphism between crystals can seriously disturb the search for good heavy-atom derivatives. It is difficult to distinguish between the real differences caused by heavy-atom substitution and the signals caused by lack of isomorphism. Even slightly non-isomorphous data sets can ultimately obscure the heavy-metal phase information. Collecting a second or third native data set can help in deciding if the non-isomorphism observed is an intrinsic characteristic of the crystals or caused by heavy-atom-soaking conditions.

(ii) For the most effective application of *DM* methods, it is important that data sets be complete, particularly in the low-resolution shells. This is because density-modification procedures rely on the relationships among all reflections; absent reflections, especially strong low-resolution ones, render the method less effective and the resultant maps less interpretable.

(iii) The reliability of the phase-probability distribution is important both for density-modification and refinement procedures which incorporate phases. Many calculations in structure determination are rather unstable and small modifications can lead to quite different behaviour. This is important because small changes can make uninterpretable maps interpretable. This problem is exacerbated when the derivative sites are correlated. This can lead to a misleadingly overestimated figure of merit, which does not necessarily result in a better electron-density map.

(iv) The maximum-likelihood refinement of a partial structure, using the experimental phases, can automatically discriminate the correct from the incorrect parts of electron-density maps and give properly weighted combined phases which are better than either the experimental or the calculated ones.

(v) Multi-crystal averaging, even between closely related forms, can improve phases dramatically when the transformation matrices are known accurately. Determining these transformations requires either some phase information for all the crystal forms to allow their refinement, or in our experience, a 'sufficiently good' model to be used in molecular-replacement searches.

This work was supported by grant BO4808 from the BBSRC. KHGV was supported by a HCMP Fellowship from the European Community. GNM is supported by a BBSRC Postdoctoral Fellowship accorded to *CCP*4 (grant No. BO5273) and EJD by an MRC grant (grant no. G9413078MB). The authors wish to thank staff on the beamlines at the European Molecular Biology Laboratory (EMBL), Hamburg, and the SRS Daresbury for assistance with data collection. Data collection at the EMBL Hamburg outstation was assisted by the European Union HCMP Large Installations programme (contract number CHGE-CT93-0040).

References

Abrahams, J. P. (1997). *Acta Cryst.* D**53**, 371–376.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Cowtan, K. D. (1994). *Jnt CCP4 ESF-EACBM Newslett. Protein Crystallogr.* **31**, 34–38.

Cowtan, K. D. & Main, P. (1996). *Acta Cryst.* D**52**, 43–48.

Fortelle, E. de la & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.

Kredich, N. M. (1996). *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, edited by F. C. Neidhart, R. Curtiss, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter & H. F. Umbarger, pp. 514–527. Washington DC: American Society for Microbiology.

Lamzin, V. S. & Wilson, K. S. (1993). *Acta Cryst.* D**49**, 129–147.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.

Oldfield, T. J. (1996). *Macromolecular Refinement. Proceedings of the CCP4 Study Weekend*, edited by E. J. Dodson, M. H. Moore, A. Ralph & S. Bailey, pp. 67–74. Warrington: Daresbury Laboratory.

Otwinowski, Z. (1991). *Proceedings of the CCP4 Study Weekend*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 80–88. Warrington: Daresbury Laboratory.

Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.

Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Acta Cryst.* D**54**, 1285–1294.

Schell, M. A. (1993). *Annu. Rev. Microbiol.* **47**, 597–626.

Terwilliger, T. C. & Berendzen, J. (1996). *Acta Cryst.* D**52**, 749–757.

Tyrrell, R., Davies, G. J., Wilson, K. S. & Wilkinson, A. J. (1994). *J. Mol. Biol.* **235**, 1159–1161.

Tyrrell, R., Verschueren, K. H. G., Dodson, E. J., Murshudov, G. N., Addy, C. & Wilkinson, A. J. (1997). *Structure*, **5**, 1017–1032.