Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Optimization of selenium substructures as obtained from *SHELXD*

Fabio Dall'Antonia,[a] Patrick J. Baker[b] and Thomas R. Schneider[a]*

[a]Department of Structural Chemistry, University of Göttingen, Tammannstrasse 4, 37077 Göttingen, Germany, and [b]Department of Molecular Biology and Biotechnology, University of Sheffield, Western Bank, Sheffield S10 2TN, England

Correspondence e-mail: trs@shelx.uni-ac.gwdg.de

Received 20 February 2003
Accepted 7 August 2003

Using the signal of naturally inbuilt or artificially introduced anomalous scatterers to derive initial phases in a macromolecular crystal structure determination has become routine in recent years. In the context of high-throughput crystallography in particular, MAD and SAD (multiple- and single-wavelength anomalous dispersion) methods are central tools. For both techniques, a crucial step is the determination of the substructure of anomalous scatterers; subsequent phasing procedures will profit from a substructure model that is as accurate as possible. The choice of the subset of the diffraction data to be used for the substructure determination has a strong influence on the quality of the substructure and can make the difference between success and failure. The accuracy of selenium substructures obtained using F_A values or various anomalous differences truncated to different resolutions has been investigated by comparing the sites determined by *SHELXD* with the selenium positions in the refined models. Based on the analysis, some recommendations for obtaining accurate and precise substructures are derived.

1. Introduction

Recent advances in X-ray sources, cryocrystallography and detector technology have enabled protein crystallographers to make use of the often very weak signal from anomalously scattering atoms for the phasing of macromolecular crystal structures. MAD phasing using SeMet-substituted protein (Hendrickson, 1991; Doublié, 1997) has become a routine procedure that allows the tackling of ever larger problems (see, for example, KPMHT; van Delft & Blundell, 2003). Phasing based on anomalous data collected at a single wavelength (SAD), although shown to be experimentally feasible by the structure determination of crambin (Hendrickson & Teeter, 1981) and underpinned theoretically by Wang (1985) more than 15 years ago, has only found widespread application very recently. After Dauter and coworkers showed that structures of the size of lysozyme and larger can be solved based on the naturally present S atoms (Dauter *et al.*, 1999) or on halide atoms introduced into the crystal by quick-soaking techniques (Dauter & Dauter, 1999), a constant stream of reports where weaker and weaker anomalous signals have been used for phasing has set in (for a recent review, see Dauter, 2002*a*).

For both MAD and SAD phasing, the determination of the substructure of the anomalous scatterers alone is a crucial step in the phasing process. To solve the substructure, first the structure factors that represent the substructure by itself need

to be prepared. In the SAD case, the anomalous differences or ΔF values calculated between reflections with indices hkl and $\overline{hkl}$ can be used for this purpose. However, such ΔF values only represent lower-limit estimates of the structure-factor amplitudes of the anomalous scatterers (Drenth, 1994). If diffraction data have been measured at several wavelengths, estimates for the full structure-factor amplitudes of the anomalous scatterers, the so-called F_A values, can be derived (Hendrickson *et al.*, 1985). A number of programs are available for the estimation of substructure structure factors, *e.g.* *CNS* (Brünger *et al.*, 1998), *DREAR* (Blessing & Smith, 1999), *MADSYS* (Hendrickson, 1991), *REVISE* (Fan *et al.*, 1993), *SOLVE* (Terwilliger, 1994) and *XPREP* (Bruker AXS, Madison, USA). In a second step, the resulting F_A or ΔF values are used as input to programs that determine the substructure by means of Patterson or direct methods or a combination thereof. Such programs include *SOLVE* (Terwilliger, 1994), *CNS* (Brünger *et al.*, 1998), *SNB* (Weeks & Miller, 1999) and *SHELXD* (Scheider & Sheldrick, 2002). Interestingly, the latter two were initially intended for the *ab initio* solution of large small-molecule structures (Usón & Sheldrick, 1999) using data to atomic resolution, but now play an important role in the field of macromolecular crystallography at lower resolution.

Naturally, a complete and precise substructure will result in better phase estimates than an incomplete and/or imprecise substructure. To this end, sophisticated methods such as the *SHARP* framework (de La Fortelle & Bricogne, 1997) have been developed to refine and complete substructures in order to derive the best possible starting phases for the respective protein structure. However, in recent experiments we found that provided high-quality diffraction data are available, substructures obtained from *SHELXD* against suitable substructure structure factors can be sufficient for successful phasing without any additional refinement or updating of the sites.

To obtain the best possible substructure from a given set of diffraction data, the choice of the type of structure factor (F_A *versus* ΔF) and of the resolution cutoff are important parameters. It has been shown previously that including data to too high resolution, for example, can be detrimental to the substructure solution process to the extent that the structure cannot be solved [*e.g.* the case of acyltransferase in Schneider & Sheldrick (2002), where the inclusion of data to higher than 3.5 Å makes it impossible to solve the substructure].

In this paper, we investigate the effect of using F_A or different anomalous difference data and of truncating the data at different high-resolution limits on the quality of the substructure. For three crystal structures where a model of the SeMet-substituted protein refined to high resolution is available, the sites found by *SHELXD* are compared with the refined positions of the respective Se atoms. The comparisons are performed using a newly developed stand-alone computer program, *SITCOM*, that allows comparison of substructures taking origin shifts, different enantiomers and symmetry operations into account. Based on the results, some recommendations for the optimum use of *SHELXD* are formulated.

Table 1
Content of the crystallographic unit cell for the test cases.

The number of residues and Se sites per asymmetric unit (AU) are listed. For the Se sites, both the expected number of sites (Exp.) and the number of Se atoms present in the refined structure (Ref.) are shown. *r/s* denotes the number of residues per Se atom and SC denotes the solvent content estimated from the refined model. d_{min} and R are the maximum resolution and the crystallographic R value for the respective model as deposited in the Protein Data Bank.

	PDB code	Residues per AU	Se per AU: Exp.	Se per AU: Ref.	*r/s*	SC (%)	d_{min} (Å)	R (%)
MODE	1b9m	2 × 265 = 530	2 × 3	6	88	59	1.75	23.4
CYAN	1dw9	10 × 156 = 1560	10 × 4	40	39	49	1.65	15.0
THDI	1f8g	4 × 384 = 1536	4 × 15	58	27	43	2.00	21.0

2. Test data

Three cases of SeMet-substituted proteins for which a refined model of the SeMet form is available were selected: molybdate-dependent transcriptional regulator (MODE; space group $P2_12_12$; a = 81.61, b = 127.24, c = 62.99 Å, $\alpha = \beta = \gamma = 90.0^\circ$; Hall *et al.*, 1999), cyanase (CYAN; space group $P1$; a = 76.34, b = 81.03, c = 82.30 Å, α = 70.3, β = 72.2, γ = 66.40°; Walsh *et al.*, 1999) and the dI component of transhydrogenase (THDI; space group $P2_1$; a = 65.9, b = 116.6, c = 102.0 Å, $\alpha = \gamma = 90.0$, $\beta = 104.2^\circ$; Buckley *et al.*, 2000). Data concerning the unit-cell contents and the quality of the refined model are summarized in Table 1. Statistics for the diffraction data are given in Table 2.

To derive reference phases for the phase comparisons, the model of THDI as obtained from the Protein Data Bank was translated into *SHELX* format using *SHELXPRO* (Sheldrick & Schneider, 1997) and the overall scale factor and two bulk-solvent parameters were refined ('`BLOC 0`' command in *SHELXL*) with *SHELXL* (Sheldrick & Schneider, 1997) for ten cycles against the high-energy remote (HRM) data.

2.1. Data analysis

All data were originally processed with *DENZO* and *SCALEPACK* (Otwinowski & Minor, 1997); details can be found in the original publications. In all cases, data were scaled independently for each wavelength. The program *XPREP* (Bruker AXS, Madison, USA) was used for the analysis of the multi-wavelength data and to derive F_A values and anomalous differences. For MODE and THDI, the analysis was based on scaled but unmerged data; for CYAN, merged data were used. During *XPREP* analysis, all data were kept at all times (*i.e.* no resolution cutoff was applied at any stage) and default settings were used. For the determination of F_A values, the f' and f'' values were refined for one cycle for each wavelength. As the resolution limits for the data to be employed for substructure determination can be chosen later on from within *SHELXD*, substructure structure factors for the full resolution range of the measured data were written to file. Two quality indicators, the signal-to-noise ratio for the anomalous differences $[\Delta F/\sigma(\Delta F)]$ and the correlation coefficient between anomalous differences measured at two wavelengths i and j,

Table 2
Diffraction data statistics for the test data.

Values for the wavelengths used for data collection were taken from the original publications. HRM, high-energy remote; PK, peak; IP, inflection point. f' and f'' are the refined values of the anomalous contributions to the scattering factor as obtained from *XPREP*. Hi defines the high-resolution bin for the statistics shown in parentheses where appropriate and Red and Cpl stand for the redundancy and completeness of the data (Friedel pairs kept separate), respectively. I and $\sigma(I)$ are the mean diffraction intensity and its standard deviation. $R_{\rm int} = \sum |\langle I\rangle - I| / \sum I$. For CYAN, only merged data were available and the statistics refer to data where Friedel pairs have been merged; the values for redundancy and $R_{\rm int}$ were taken from Walsh *et al.* (1999).

	Wavelength	λ (Å)	f'	f''	Hi	Red	Cpl (%)	$I/\sigma(I)$	$R_{\rm int}$ (%)
MODE	HRM	0.8855	−2.7	2.4	2.7–2.6	3.4 (1.9)	97.1 (96.5)	19.9 (8.6)	4.0 (10.3)
	PK	0.9782	−6.6	6.5	2.7–2.6	4.8 (2.7)	94.9 (78.6)	19.0 (5.2)	3.8 (12.4)
	IP	0.9779	−2.8	2.7	2.7–2.6	3.0 (1.8)	91.8 (62.1)	19.3 (5.5)	4.1 (17.8)
CYAN	HRM	0.94645	−1.5	2.3	2.5–2.4	*3.9*	96.2 (95.0)	23.9 (20.6)	*2.3 (2.9)*
	PK	0.97933	−6.3	5.2	2.5–2.4	*3.3*	94.0 (83.2)	22.8 (17.9)	*5.9 (7.9)*
	IP	0.97947	−7.0	3.4	2.5–2.4	*3.8*	96.3 (85.0)	20.4 (14.1)	*4.5 (6.4)*
	LRM	1.07813	−2.4	0.5	2.5–2.4	*3.8*	85.0 (40.9)	24.6 (20.2)	*2.7 (4.2)*
THDI	HRM	0.9686	−4.1	3.6	2.1–2.0	2.5 (2.4)	89.9 (83.4)	8.9 (2.2)	4.3 (25.6)
	PK	0.9794	−8.9	6.4	2.1–2.0	2.3 (2.1)	81.8 (71.5)	10.8 (2.4)	4.7 (27.6)
	IP	0.9796	−9.9	3.0	2.1–2.0	2.3 (2.2)	81.7 (71.6)	10.9 (2.6)	4.4 (25.4)

CC(ΔF_i, ΔF_j) (Schneider & Sheldrick, 2002), both averaged in resolution bins by *XPREP*, were inspected (Fig. 1).

2.2. Substructure determination

Selenium substructures were determined by running *SHELXD* (Schneider & Sheldrick, 2002) against ΔF or F_A values using default parameters. Patterson seeding was used to generate initial phases for the substructures; after termination of the dual-space recycling part, which only uses the reflections with large normalized structure factors, the occupancies of the sites were refined against the complete set of substructure structure factors (also including the weak reflections). The `SHEL` keyword in *SHELXD* was used to completely exclude data outside a given resolution range from the substructure-determination process. The number of trials was limited to 100.

2.3. Phasing calculations and model building

For THDI, phases were determined for different substructures using *SHELXE* (Schneider & Sheldrick, 2002) employing the HRM data as native. The lists of substructure sites were taken as provided by *SHELXD* without any editing. The solvent content was estimated to be 43% from the number of ordered residues in the final model, assuming a volume of 140 Å^3 per residue. Ten cycles of density modification were run. Comparison of phase sets was performed using the method of Lunin & Woolfson (1993) as implemented in a new prerelease version of *SHELXPRO* (Sheldrick & Schneider, 1997).

Automatic model building was performed with *ARP*/*wARP* version 6.0 (Perrakis *et al.*, 1999) and *FFFEAR* (Cowtan, 1998) employing α-helices as search fragments. For both programs, standard parameters as provided by the *CCP*4 graphical user interface (Collaborative Computational Project, Number 4, 1994) were used.

2.4. Analysis of substructures

The comparison of substructure sites with the refined atomic positions is complicated by the fact that in order to measure the respective distances, the substructure sites first have to be moved to the same asymmetric unit, the same origin and the same enantiomorph as the refined structure. A computer program, *SITCOM*, has been written to automatize such comparisons. For non-*P*1 cases, *SITCOM* uses essentialy the same approach as the recently described program *NANTMRF* (Smith, 2002). In addition to the functionality available in *NANTMRF*, *SITCOM* also provides facilities for comparisons in space group *P*1.

For non-polar space groups, *SITCOM* transforms both the original list of sites and its enantiomer by application of all symmetry operators and origin shifts belonging to the respective space group. For each combination of geometrical transformations, the number of hits, where a hit is defined as a situation where a refined atom can be found within a distance of 2.0 Å from a site or its symmetry-related copy, is recorded. For the transformation with the largest number of hits, the mean distance between sites and refined atoms $\langle d\rangle$ is calculated.

For polar space groups, a similar strategy is used. In a first step, corresponding site atom pairs are identified by scoring hits based on a two-dimensional distance criterion (here 2.0 Å), which effectively compares the sites and the atoms in a projection onto a plane perpendicular to the polar axis. Once pairs of sites and atoms have been identified, the shift along the polar axis is determined in an iterative fashion. In each round of the iteration, a mean distance along the polar direction between pairs (as identified by the two-dimensional distance criterion) of sites and atoms is first determined. The pairs with distances strongly deviating from the mean distance are then discarded and a new mean value for the shift along the polar axis is calculated. Finally, the corresponding shift is applied to the entire list of sites. This iterative scheme usually converges in less than five cycles.

In the triclinic cases, a three-step method is applied to both enantiomers to find a three-dimensional translation vector

that is common to as many pairs of sites and refined atoms as possible. First, a systematic search is performed to find two pairs of sites and atoms that have parallel connecting vectors. The corresponding shift is then applied to all sites, taking neighbouring unit cells into account. After all correspondences between refined atoms and shifted sites are established, the fit between sites and atoms is optimized by superimposing the centres of mass of the two constellations.

3. Results and discussion

3.1. Quality of the test data

All test data sets are of high quality and enabled the corresponding crystal structures to be solved. MODE and THDI show the expected behaviour of the resolution-dependent quality indicators $\Delta F/\sigma(\Delta F)$ and $CC(\Delta F_i, \Delta F_j)$ (Fig. 1): as it becomes increasingly difficult to measure accurately anomalous differences when going from stronger reflections at low resolution to weaker reflections at high resolution, both the signal to noise, $\Delta F/\sigma(\Delta F)$, and the correlation coefficient between signed anomalous differences, $CC(\Delta F_i, \Delta F_j)$, decrease. Overall, the signal to noise is somewhat lower for MODE than for THDI. This is most likely to be a consequence of the different proportion of Se atoms with respect to the unit-cell content: MODE contains about three times as many residues per Se atom than THDI (88 *versus* 27 residues; Table 1).

The CYAN data are clearly exceptional. $\Delta F/\sigma(\Delta F)$ exhibits only a small resolution-dependence: the lower quality of the peak (PK) data between 3.4 and 4.0 Å is probably owing to the increased X-ray background caused by the so-called water ring in this region of reciprocal space. $CC(\Delta F_i, \Delta F_j)$ is essentially independent of resolution.

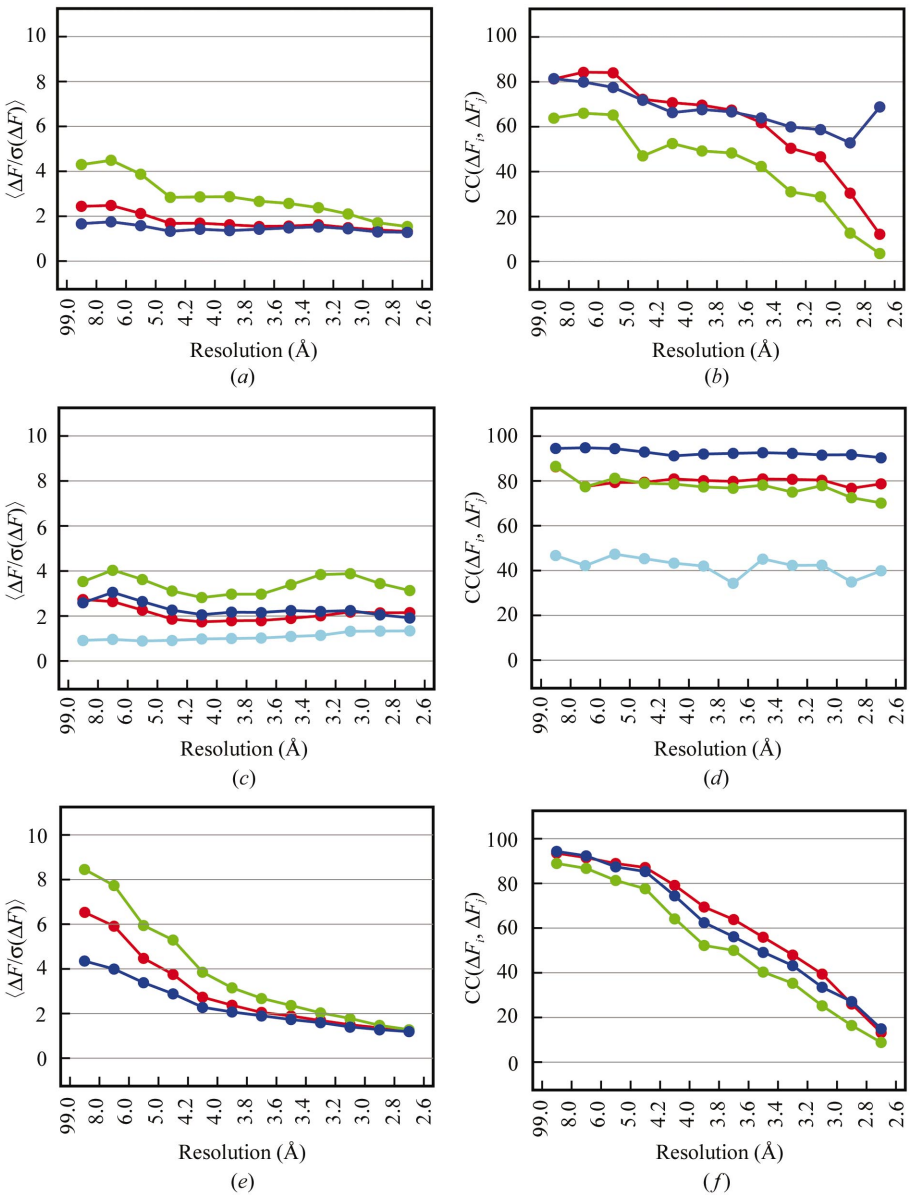

Figure 1
Resolution-dependent quality indicators for the anomalous differences for MODE (*a*, *b*), CYAN (*c*, *d*) and THDI (*e*, *f*). (*a*), (*c*) and (*e*) show $\langle\Delta F/\sigma(\Delta F)\rangle$ in resolution bins for PK (green), HRM (red), IP (blue) and LRM (cyan); (*b*), (*d*) and (*f*) show $CC(\Delta F_i, \Delta F_j)$ in resolution bins: red = $CC(\Delta F_{HRM}, \Delta F_{PK})$; green = $CC(\Delta F_{HRM}, \Delta F_{IP})$; blue = $CC(\Delta F_{PK}, \Delta F_{IP})$; cyan = $CC(\Delta F_{HRM}, \Delta F_{IP})$.

3.2. Molybdate-dependent transcriptional regulator

The Se substructure of molybdate-dependent enhancement factor can be solved by taking many different routes (Table 3). For all scenarios evaluated here, apart from those based on the F_{IP} (inflection point) data, the complete substructure consisting of six Se sites is readily found. For the IP data, five out of the six sites are found unless the data between 3.0 and 2.6 Å are included. The missing site always corresponds to the Se atom with the highest B value in the refined model (Se_{A1} with B = 57.3 Å^2).

For all four sets of substructure structure factors investigated, the best substructure in terms of mean distance between sites and refined positions of the corresponding atoms is obtained when the data are truncated at 3.0 Å resolution, supporting the previous suggestion (Schneider & Sheldrick, 2002) to limit substructure structure factors to the resolution where $CC(\Delta F_i, \Delta F_j)$ drops below ∼30% (Fig. 1*b*).

For the solutions obtained from F_A values, F_{PK} and F_{HRM} differences, the mean distance between sites and refined atoms is between 0.16 and 0.18 Å, a difference which is somewhat

Table 3
Substructure solution for *MODE*.

For different resolution cutoffs d_{min} and substructure structure-factor sets, # denotes the number of successful trials (delivering six sites within 2.0 Å of the the corresponding refined atom) per 100 starting phase sets, CC_1 is the highest correlation coefficient between E_{obs} and E_{calc} as defined by Fujinaga & Read (1987) and $\langle d \rangle$ is the mean distance between the sites in solution with CC_1 and the respective Se atoms in the refined structure as determined by *SITCOM*. For the best substructure in each column, figures are printed in bold.

	F_A			F_{PK}			F_{HRM}			F_{IP}		
d_{min} (Å)	#	CC_1 (%)	$\langle d \rangle$ (Å)	#	CC_1 (%)	$\langle d \rangle$ (Å)	#	CC_1 (%)	$\langle d \rangle$ (Å)	#	CC_1 (%)	$\langle d \rangle$ (Å)
2.6	96	54.8	0.22	99	44.2	0.21	97	27.4	0.21	0	7.9	n/a
3.0	100	**68.4**	**0.18**	98	**50.4**	**0.16**	98	**34.4**	**0.18**	51	**22.8**	**(0.25)**
3.5	100	65.5	0.19	100	54.8	0.19	88	41.0	0.22	46	28.8	(0.32)
4.0	100	78.6	0.25	100	57.2	0.25	85	46.5	0.26	61	33.4	(0.36)

Table 4
Substructure solution for CYAN and THDI.

For different resolution cutoffs d_{min} and substructure structure-factor sets, CC_1 is the highest $CC(E_{obs}, E_{calc})$ obtained for a run of *SHELXD* and # denotes the number of sites in the substructure with CC_1 that are closer than 2.0 Å to the respective refined atom. $\langle d \rangle$ is the mean distance between the sites of the best solution and the respective Se atoms in the refined structure as determined by *SITCOM*. For the best substructure in each column, figures are printed in bold.

	F_A			F_{pk}			F_{hrm}			F_{ip}		
d_{min} (Å)	CC_1 (%)	#	$\langle d \rangle$ (Å)	CC_1 (%)	#	$\langle d \rangle$ (Å)	CC_1 (%)	#	$\langle d \rangle$ (Å)	CC_1 (%)	#	$\langle d \rangle$ (Å)
2.4	**60.1**	**40**	**0.24**	**50.9**	**39**	**0.26**	**48.9**	**40**	**0.20**	**49.5**	**40**	**0.25**
3.0	63.6	40	0.25	52.3	38	0.30	49.8	40	0.26	51.8	39	0.29
3.5	64.4	40	0.29	51.7	38	0.32	49.3	40	0.32	51.9	39	0.32
4.0	63.3	39	0.45	51.1	38	0.47	48.7	40	0.47	51.4	40	0.45
2.0	35.9	52	0.44	37.4	54	0.29	26.4	54	0.43	11.6	0	n/a
2.5	55.1	56	0.36	**46.4**	**57**	**0.27**	37.8	55	0.32	17.2	0	n/a
3.0	66.9	58	0.36	51.3	55	0.32	**44.7**	**57**	**0.37**	38.1	49	0.56
3.5	**71.7**	**58**	**0.35**	53.9	56	0.39	49.2	57	0.40	**42.1**	**52**	**0.63**
4.0	74.4	58	0.42	32.5	0	n/a	29.5	0	n/a	38.9	37	0.76

smaller than the coordinate uncertainty that would be expected for a structure refined at the 'native' resolution of the data, 2.6 Å.

When substructures determined using the same resolution cutoff but different substructure factors are compared, the quality of the substructures is very similar. There is a slight tendency for the sites determined from the PK data to be marginally more accurate than the corresponding sites from the HRM data. This may arise from the signal to noise being higher for the F_{PK} than for the F_{HRM} data, which in turn could be a consequence of the higher redundancy of the data collected at the PK wavelength (Table 2).

The single-wavelength quality indicator (Fig. 1*a*) does not predict the very different behaviour of the HRM and the IP data: both data sets show an almost identical behaviour of $\Delta F/\sigma(\Delta F)$ against resolution, but the substructures obtained from the IP data are not complete and are much less precise than those obtained from the HRM data.

The multi-wavelength quality indicator $CC(\Delta F_i, \Delta F_j)$ shows a lower correlation between the F_{HRM} and F_{IP} data (green line in Fig. 1*b*) than for the other other two combinations of wavelengths. At first sight this does not directly indicate a lower quality of the IP data, as the correlation between F_{PK} and the F_{IP} (blue line in Fig. 1*b*) is very similar to that between F_{PK} and the F_{HRM} (red line in Fig. 1*b*). However, if one takes into account that a standard linear correlation coefficient such as that used here is not sensitive if the data under comparison are suffering from similar systematic errors, the high correlation between the PK and the IP data could well be an artefact. In fact, the wavelengths with which these two data sets were collected differ by only 0.0003 Å (Table 2), making it likely that the systematic errors in the two data sets are at least more related than for other pairs of data sets. In addition, this argument could also explain the rather unphysical observation of an increase in $CC(\Delta F_{PK}, \Delta F_{HRM})$ for the highest resolution bin in the blue curve in Fig. 1(*b*).

The correlation coefficient between E_{obs} and E_{calc} for a successful substructure solution is in general higher for the F_A-based than for the ΔF-based experiments, reflecting the fact that F_A values are true estimates of the structure factors of the Se atoms alone, whereas the ΔF values represent only lower limit estimates.

3.3. Cyanase

For the very high quality data for cyanase, the inclusion of higher resolution data into the substructure-determination process is always beneficial (Table 4). When data to the maximum resolution of 2.4 Å are used, the complete substructure of 40 Se sites is always found, except for the case of the anomalous differences derived from the PK data; here, one site is missing. The mean distance between substructure sites and refined atoms is between 0.20 and 0.26 Å for all data sets, which is of the order of the coordinate error expected for a refined structure at 2.4 Å resolution. Interestingly, even when the data are truncated to 4.0 Å complete or almost complete substructures with mean coordinate errors of less than 0.5 Å are found by *SHELXD*.

Generally, the results obtained from F_A values and the anomalous differences determined at PK, HRM and IP wavelengths are very comparable. $CC(E_{obs}, E_{calc})$ values for the best solution are very similar for all ΔF-based experiments and are systematically higher for the substructures determined against F_A values.

For the F_{LRM} (low-energy remote) data, no sites could be found for any subset of the structure factors (data not shown). This shows that although the very small anomalous signal present at the LRM wavelength had been measured accurately [$CC(F_{HRM}, F_{LRM})$ is around 40% for all resolution bins; Fig. 1], the derived ΔF values are not sufficiently precise to furnish a solution of the Se substructure.

3.4. Transhydrogenase dI

For the case of the dI component of transhydrogenase B, the most complete substructures are obtained using F_A values

truncated between 3.0 and 4.0 Å resolution. In all these cases, out of the 59 sites present in the refined structure, only the Se atom of Met*D*226 is not found. However, with a refined *B* value of 82.0 Å^2, this atom is also not very well defined in the final model. For the 58-site solutions, the mean distance between the found and the refined sites of 0.35–0.36 Å is of the order that would be expected for the comparison of pairs of atoms in two independently refined structures at 3 Å resolution. In this case, limiting the F_A values to the resolution where CC(ΔF_i, ΔF_j) drops below 30%, *i.e.* ∼2.5 Å, would not have given the very best solution. The respective result is nevertheless still acceptable (56 sites with $\langle d \rangle$ = 0.36 Å); the less than expected quality may be related to the incompleteness of the reflection data (Table 2).

When anomalous differences are employed, the resolution limit at which the most complete substructures (57 sites for PK and HRM, 52 sites for IP) appear varies with the quality of the data. The higher the signal to noise (Fig. 1*e*), the more data can be profitably used in the substructure-determination process. Also, although the substructures are relatively complete for all three cases, there is a notable deterioration in the precision of the positions from HRM to IP: $\langle d \rangle$ increases from 0.27 to 0.63 Å. For all resolution cutoffs, the data collected at the inflection point of the f'' curve consistently yield the worst substructures both in terms of completeness and coordinate precision.

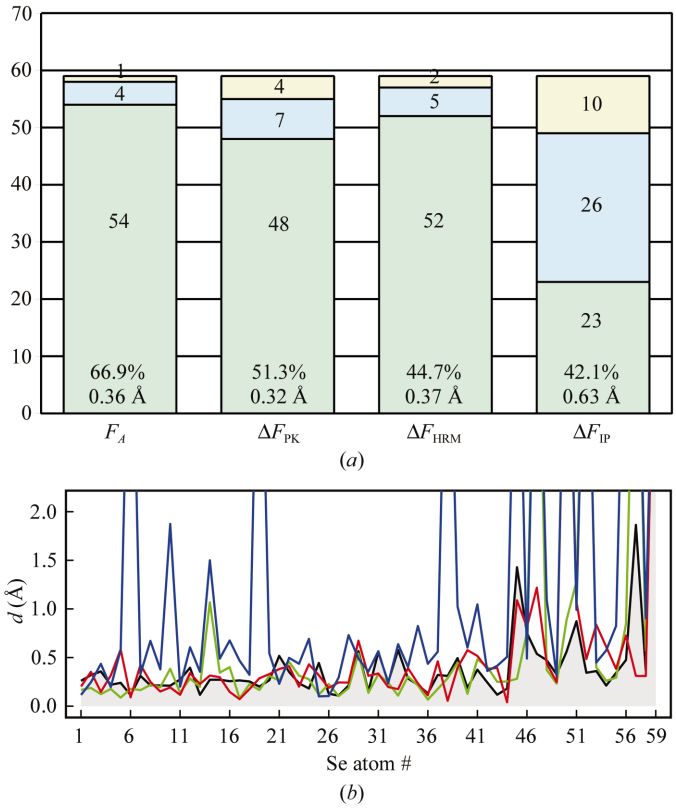

Figure 2
Analysis of substructure sites for *SHELXD* runs against F_A values or different anomalous differences with data truncated to 3 Å resolution. (*a*) The total length of the bars corresponds to the number of Se atoms in the refined structure (59). The green parts represent the part of the site list that is continous; for example, for the best solution against F_A data, site number 55 is the first incorrect one. The blue parts stand for the number of sites that can be found in the 'discontinous' part of the list of 84 sites output by *SHELXD* when 60 sites have been requested. The yellow parts correspond to the missing sites. (*b*) Distance between sites and refined position of the corresponding Se atom against number of the Se atom, in which the 59 refined Se atoms have been sorted in order of increasing *B* values; F_A values, black; F_{PK}, green; F_{HRM}, red; F_{IP}, blue.

Table 5
Phasing of THDI based on different substructures.

Results for *SHELXE* phasing based on sites obtained from different sources (column 'Data') including refined atoms taken from the final model in combination with F_A values as provided by *XPREP*. For each substructure, the number of sites #, the mean distance between sites and refined Se atoms $\langle d \rangle$ and the mean phase error $\langle \Delta\varphi \rangle$ between the phases determined by *SHELXE* and the phases calculated from the refined model are given.

Data	#	$\langle d \rangle$ (Å)	$\langle \Delta\varphi \rangle$ (°)
Ref	59	0.00	30.0
F_A	58	0.36	32.4
F_{PK}	55	0.32	33.6
ΔF_{HRM}	57	0.37	34.0
ΔF_{IP}	49	0.56	40.5

A more detailed analysis of the site lists obtained using different substructure structure factors, all limited to 3.0 Å, is shown in Fig. 2. The figure shows that the sites lists obtained from F_A, F_{PK} and F_{HRM} are very comparable, although CC(E_{obs}, E_{calc}) varies between 66.9 and 44.7%. For Se atoms #1 to #40, where Se atom #40 has a *B* value of 23.3 Å^2, the corresponding sites found by *SHELXD* from the F_A, F_{PK} or the F_{HRM} data are mostly closer than 0.5 Å. The only significant exception is Se atom #14, for which the site found against the HRM data is more than 1 Å away. For these three cases, the number of sites not found in the initial continous part of the site list is relatively small (4, 5 and 7; Fig. 2*a*). Given that the non-crystallographic symmetry operators can be established firmly based on the 40 or so strong sites, the sites in the discontinous part of a solution can most likely be identified by checking their consistency with the non-crystallographic symmetry.

The substructure determined against the IP data is of much lower quality and it is not clear whether the 26 sites in the discountinous part of the site list could have been identified by non-crystallographic symmetry given that only the first 23 sites form a continous set. Also, the missing or unprecisely positioned sites correspond not only to refined atoms with high *B* values, but a number of Se atoms with relatively low *B* value (#6, #10, #14 and #19) are only located with large error or not found at all.

The effect of the quality of the substructure on the resulting crystallographic phases has been investigated using the various sets of sites obtained from *SHELXD* runs with data truncated to 3 Å as input to the phasing program *SHELXE* (Sheldrick, 2002). For the phase calculation, the list of sites provided by *SHELXD* was not edited, *i.e.* 84 sites with varying occupanies were directly submitted to *SHELXE*. As expected, the more complete and the more precise a substructure, the smaller the phase error resulting from the phasing calculation is (Table 5). The high quality of the substructures derived from

the F_A and the F_{PK} data can be appreciated if the phases obtained using these substructures are compared with the phases originating from using the refined sites: the mean phase error is only 3° larger for the former than for the latter scenario.

Although for this case the observables-to-parameters ratio of 1.7 is rather low, the electron-density maps obtained from *SHELXE* based on the F_A or F_{PK} sites can be readily automatically interpreted. For example, for the F_A case, *ARP*/*wARP* places 1280 amino acids in 50 cycles.

For less good phase sets obtained from various combinations of substructures and substructure structure factors, automatic map interpretation becomes progressively less straightforward: the number of residues placed in a given number of *ARP*/*wARP* cycles decreases (data not shown). At a phase error around 40°, using *ARP*/*wARP* with standard parameters ceases to work. However, for example, the electron-density map obtained from the F_{IP}-derived sites with a mean phase error of 40.5° is still of sufficient quality to support the automatic positioning of α-helical fragments with *FFFEAR* (data not shown).

4. Conclusions

4.1. Measuring data quality

Provided accurate estimates of the uncertainties in the measured diffraction data are available, the signal to noise for the anomalous differences, $\langle \Delta F/\sigma(\Delta F)\rangle$, is an acceptable indicator of the *precision* of the measured data. Unfortunately, when plotted against resolution, the region in which the decision about truncating the data has to be made is rather flat.

The correlation coefficient between signed anomalous differences measured at two different wavelengths i and j, $\mathrm{CC}(\Delta F_i, \Delta F_j)$, is a useful measure of the *accuracy* of the data. When using this measure, however, one should keep in mind how the experiment was performed. For the case of a MAD experiment, the data collected at the high-energy remote wavelength can be most reliably used as a reference, as from an experimental point of view these are the data that are the least difficult to collect accurately (f'' is significant and the f'' curve is flat). A high value of $\mathrm{CC}(\Delta F_i, \Delta F_j)$ can sometimes be misleading, as a linear correlation coefficient is not sensitive to identical systematic errors in the i and j measurements. Such identical or related systematic errors can arise for data sets that are collected at very similar wavelength, which is often the case for the IP and the PK data, or if the data sets compared suffer from strong background in the same regions of reciprocal space, for example owing to the presence of pronounced ice or water rings.

4.2. Choosing data for substructure solution

The previous suggestion to truncate substructure data for MAD phasing at the resolution where the correlation between signed anomalous differences drops below 30% has been substantiated. For both MODE and THDI, the best or close to the best substructues are obtained following this suggestion. At the suggested resolution cutoff the advantage of including more data, thus improving the data-to-parameter ratio, is outweighed by the disturbances introduced by the inclusion of inaccurately measured data. For data of outstanding quality such as the CYAN data, the more data are included into the substructure determination, the higher the quality of the substructure will be.

In the framework of methods used here, the use of F_A values is marginally advantageous over the use of anomalous differences obtained from the PK data. However, in practice the differences are probably negligible.

Not surprisingly, the anomalous differences derived from the IP data are the worst for all the cases discussed. This illustrates the difficulty of measuring accurate anomalous differences in the steep region around the inflection point of the f'' curve.

4.3. Measuring the quality of substructure solutions

The overall quality of a substructure solution can be measured by the correlation coefficient between observed and calculated E values, $\mathrm{CC}(E_{obs}, E_{calc})$. Generally, a higher value of this figure of merit indicates a better substructure. However, the absolute magnitude of $\mathrm{CC}(E_{obs}, E_{calc})$ for the best substructure that can be obtained against a given set of structure factors strongly depends on the quality of these structure factors. F_A-value-based substructure determinations will normally yield higher correlation coefficients than substructures determined against anomalous differences, reflecting the fact that F_A values are more accurate estimates of the true structure factors of the anomalous scatterers than are ΔF values. Furthermore, inclusion of weaker data (either by using a data set containing generally weaker data or by including higher resolution data) will produce smaller correlation coefficients for correct solutions.

In the present study, several cases where Se-substructures with $\mathrm{CC}(E_{obs}, E_{calc})$ of less than 30% represented correct (albeit sometimes only partially complete) solutions have been found. High correlation coefficients are not always a sufficient condition for the correctness of a substructure; *e.g.* for F_A values for THDI truncated to 3.5 Å, some site lists with $\mathrm{CC}(E_{obs}, E_{calc})$ of more than 50% appeared that did not have a single site at less than 2 Å of the refined structure; in this case, the correct solutions had correlation coefficients of more than 70% (Table 4).

In situations where the sole use of $\mathrm{CC}(E_{obs}, E_{calc})$ does not allow a clear decision, other criteria for the correctness of the substructure such as the consistency of the solution with the Patterson or the presence of non-crystallographic symmetry should be evaluated. In *SHELXD*, both these criteria can be conveniently checked in the Patterson crossword table (Sheldrick *et al.*, 1993).

4.4. Future perspectives

Even for a relatively large structure such as THDI, whose solution was still a great achievement in 1999, the timescales

for structure solution have reduced dramatically. Starting from scaled data at different wavelengths, a procedure (all programs run with default parameters on a 2 GHz Linux PC) using *XPREP* for data analysis (5 min), *SHELXD* for substructure determination (6 min) and *SHELXE* for phase calcuation (4 min) produces an electron-density map in 15 min that can be readily interpreted by *ARP*/*wARP*. This definitely opens the possibility of solving a structure while the data are still being collected, underpinning approaches such as the recently suggested 1.5-wavelength method (Dauter, 2002*b*). Nevertheless, improved statistical measures for data quality would be useful to aid decision-making in cases where data quality cannot be evaluated by online structure solution.

As computers become still faster, reducing the timescales for structure solution in straightforward cases even more, one could consider redetermining structures from the original data whenever necessary or when new technology becomes available. However, an absolute prerequisite of this approach is the availability of the respective experimental data in databases (Jiang *et al.*, 1999).

We thank Martin Walsh, Bill Hunter and Gordon Leonard for providing the original MAD data for cyanase and molybdate-dependent enhancement factor. We are grateful to George M. Sheldrick for discussions and advice. This work was supported by the European Union (QLRI-CT-2000-00398).

References

Blessing, R. H. & Smith, G. D. (1999). *J. Appl. Cryst.* **32**, 664–670.
Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Buckley, P. A., Jackson, J. B., Schneider, T. R., White, S. A., Rice, D. W. & Baker, P. J. (2000). *Structure. Fold. Des.* **8**, 809–815.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Cowtan, K. (1998). *Acta Cryst.* D**54**, 750–756.
Dauter, Z. (2002*a*). *Curr. Opin. Struct. Biol.* **12**, 674–678.
Dauter, Z. (2002*b*). *Acta Cryst.* D**58**, 1958–1967.
Dauter, Z. & Dauter, M. (1999). *J. Mol. Biol.* **289**, 93–101.
Dauter, Z., Dauter, M., de La Fortelle, E., Bricogne, G. & Sheldrick, G. M. (1999). *J. Mol. Biol.* **289**, 83–92.
Delft, F. van & Blundell, T. L. (2003). *Acta Cryst.* A**58**, C239.
Doublié, S. (1997). *Methods Enzymol.* **276**, 523–530.
Drenth, J. (1994). *Principles of Protein X-ray Crystallography*. New York: Springer–Verlag.
Fan, H.-F., Woolfson, M. & Yao, J.-X. (1993). *Proc. R. Soc. London Ser. A*, **442**, 13–32.
Fujinaga, M. & Read, R. J. (1987). *J. Appl. Cryst.* **20**, 517–521.
Hall, D. R., Gourley, D. G., Leonard, G. A., Duke, E. M. H., Anderson, L. A., Boxer, D. H. & Hunter, W. N. (1999). *EMBO J.* **18**, 1435–1446.
Hendrickson, W. A. (1991). *Science*, **254**, 51–58.
Hendrickson, W. A., Smith, J. L. & Sheriff, S. (1985). *Methods Enzymol.* **115**, 41–55.
Hendrickson, W. A. & Teeter, M. M. (1981). *Nature (London)*, **290**, 107–113.
Jiang, J., Abola, E. & Sussman, J. L. (1999). *Acta Cryst.* D**55**, 4.
La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.
Lunin, V. Y. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 530–533.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Perrakis, A., Morris, R. J. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.
Sheldrick, G. M. (2002). *Z. Kristallogr.* **217**, 644–650.
Sheldrick, G. M., Dauter, Z., Wilson, K. S., Hope, H. & Sieker, L. C. (1993). *Acta Cryst.* D**49**, 18–23.
Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol.* **277**, 319–343.
Smith, G. D. (2002). *J. Appl. Cryst.* **35**, 368–370.
Terwilliger, T. C. (1994). *Acta Cryst.* D**50**, 11–16.
Usón, I. & Sheldrick, G. (1999). *Curr. Opin. Struct. Biol.* **9**, 643–648.
Walsh, M. A., Otwinowski, Z., Perrakis, A., Anderson, P. M. & Joachimiak, A. (1999). *Structure Fold. Des.* **8**, 505–514.
Wang, B. C. (1985). *Methods Enzymol.* **115**, 90–111.
Weeks, C. M. & Miller, R. (1999). *J. Appl. Cryst.* **32**, 120–124.