Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

The application of hierarchical cluster analysis to the selection of isomorphous crystals

Rita Giordano,[a] Ricardo M. F. Leal,[a] Gleb P. Bourenkov,[b] Sean McSweeney[a] and Alexander N. Popov[a]*

[a]ESRF, 6 Rue Jules Horowitz, 38043 Grenoble, France, and [b]EMBL Hamburg Outstation, c/o DESY, Notkestrasse 85b, 22607 Hamburg, Germany

Correspondence e-mail: apopov@esrf.fr

Received 24 January 2012
Accepted 15 February 2012

It is generally assumed that the quality of X-ray diffraction data can be improved by merging data sets from several crystals. However, this effect is only valid if the data sets used are from crystals that are structurally identical. It is found that frozen macromolecular crystals very often have relatively low structure identity (and are therefore not isomorphous); thus, to obtain a real gain from multi-crystal data sets one needs to make an appropriate selection of structurally similar crystals. The application of hierarchical cluster analysis, based on the matrix of the correlation coefficient between scaled intensities, is proposed for the identification of isomorphous data sets. Multi-crystal single-wavelength anomalous dispersion data sets from four different protein molecules have been probed to test the applicability of this method. The use of hierarchical cluster analysis permitted the selection of batches of data sets which when merged together significantly improved the crystallographic indicators of the merged data and allowed solution of the structure.

1. Introduction

The flash-cooling of crystals has become commonplace in the field of macromolecular crystallography (MX) and today over 90% of all protein structural data are obtained at cryogenic temperatures (Garman, 2010). The main advantages of cryo-cooling for MX data collection are mitigation of the rate of radiation damage, improvement of the diffraction resolution reachable and the ability to carry out crystal screening and ranking in advance of data collection. Unfortunately, cryogenic techniques can also introduce structure artefacts resulting directly from temperature effects or indirectly from the addition of cryoprotectants and temperature-induced pH changes (Juers & Matthews, 2001; Halle, 2004; Dunlop *et al.*, 2005). There is also the additional complication that the flash-cooling process is not completely under our control. For instance, among other effects, the cooling rates of the crystals may be different for crystals of different sizes, the thickness of the remaining cryoprotecting liquid surrounding the crystal will vary and the velocity of crystal transfer from the cryo-buffer to the cryogen may be different. As a result, although the crystals are apparently identical at room temperature, cryocooled crystals are rarely if ever absolutely structurally identical and often have slightly different unit-cell parameters, leading to issues of non-isomorphism that have been understood since the dawn of the study of macromolecules by crystallographic methods.

doi:10.1107/S0907444912006841

Table 1
Data collection for all proteins included in this study.

Structure	Space group	Total No. of crystals	Inverse-beam strategy	X-ray energy (keV)	Resolution (Å)	Rotation range (°)	Oscillation range (°)	Dose (MGy)	No. of residues	No. of anomalous scatterers	Anomalous scatterer	Expected anomalous signal (%)
Insulin	$I2_13$	12	Yes	9	2.2	50 + 50	2	1	51	6	S	1.2
Trypsin	$P3_121$	17	No	9	2.2	360	2.25	1	223	14/1	S/Ca	0.9
Thaumatin	$P4_121$	13	Yes	6.8	2	100 + 100	1	2	207	17	S	1.6
AroF	$P6_122$	12	Yes	12.6	3.5	90 + 90	1.5	1	407	11	Se	4.5

The amount of diffraction data that can be obtained from a single crystal is limited by radiation damage (Garman, 2010; Ravelli & Garman, 2006), mainly owing to the resolution-dependent reduction in diffraction intensity. In general, fewer data can be obtained from a small crystal before significant radiation damage occurs, as the diffracting volume is lower. Specific chemical modifications occur in the protein following a relatively small dose of irradiation (less than 2 MGy), affecting the success of SAD/MAD structure solution (Cianci *et al.*, 2008). In those cases where several crystals of the same type are available, the result of the structure study can be substantially improved by using a multi-crystal data-collection strategy, as has often been applied for room-temperature data collection (Blundell & Johnson, 1976). For low-temperature data collection the inherent non-isomorphism of cryocooled crystals discussed above may make multi-crystal data collection significantly more challenging. In order to gain from the use of multiple-crystal data-collection strategies, the central task is how to identify isomorphous crystals such that appropriate partial data sets can be merged in order to obtain an improved final data set.

In this paper, the applicability of a hierarchical cluster analysis for selecting isomorphous crystals has been tested. Several multi-crystal data sets were collected with the goal of solving the structures of insulin, bovine trypsin and thaumatin using the (weak) anomalous signal from S atoms. Additional experiments were performed using crystals of selenium-labelled chorismate synthase (AroF) from *Mycobacterium tuberculosis*. In all cases, it was confirmed that a much more accurate anomalous signal and ultimately success in substructure determination was obtained by merging data from multiple crystals preselected according to the results of the cluster analysis.

2. Materials and methods

2.1. Sample preparation

Bovine Zn-free insulin (INS), thaumatin (THA) and bovine pancreatic trypsin (TRY) samples used for crystallization were purchased from Sigma. The crystals were grown and cryoprotecting solutions were prepared according to the standard procedures described in the literature. Crystallization plates with crystals of the Se derivative of AroF (PDB entry 2o11; M. Bruning, G. P. Bourenkov, N. I. Strizhov & H. D. Bartunik, unpublished work) were kindly provided by Galina Kachalova and Hans Bartunik within the framework of the European BIOXHIT programme. Relevant crystal characteristics and expected anomalous signals are summarized in Table 1.

2.2. Data collection and processing

The diffraction measurements were carried out on beamline ID23-1 at the European Synchrotron Radiation Facility (ESRF; Nurizzo *et al.*, 2006) using an ADSC Q315R CCD detector. Only very small crystals of INS, THA and AroF, with largest dimensions of between 20 and 30 µm, were selected for data collection. TRY crystal sizes were between 30 and 50 µm. The crystals were either first transferred to cryoprotecting solution (for INS, THA and TRY) or mounted directly (AroF) from the crystallization solution using large nylon loops or litho-meshes. The mounts (loops) containing multiple sample were cryocooled in a 100 K Oxford Cryostream 600 nitrogen-gas stream.

For planning of the diffraction measurements, the program *BEST* (Bourenkov & Popov, 2010) was used. Identical data-collection conditions (resolution limit, exposure time, rotation range and oscillation width; Table 1) were applied to each type of sample. The absorption dose rates were estimated using the *RADDOSE* program (Paithankar *et al.*, 2009). The incident-beam intensity was attenuated to provide a total absorption dose per data set of less than 2 MGy in all cases; this implies that no significant radiation damage is expected for any of the data sets. For small crystals of INS, THA and AroF two rotation ranges separated by 180° were chosen to ensure accurate sample centring in the beam during both sweeps (*i.e.* avoiding the orientations in which the crystal image in the on-axis microscope would be occluded by the mounting loop) and to ensure that only one crystal is being hit by the beam at any time.

Data were indexed, integrated and scaled using *XDS*/*XSCALE* (Kabsch, 1993, 2010*a*,*b*). In order to ensure consistent indexing, one data set was arbitrarily chosen as the reference data set in the input file of *XDS* for each of the four samples. Standard data-processing statistics are summarized in Tables 2, 3, 4 and 5. The data sets are denoted L*X*_*Y*, where *X* enumerates the mounting loop and *Y* the crystal in the loop.

2.3. Clustering of diffraction data

The most popular clustering algorithm in biological applications and crystallography is hierarchical cluster analysis (Barr *et al.*, 2004; Buehler *et al.*, 2009; Hofmann *et al.*, 2009). It traditionally represents the hierarchy as a tree, or a dendrogram, with individual elements at one end and a single cluster

Table 2
Insulin: statistics for individual data sets, for the cluster and for all data sets merged.

Values in parentheses are for the highest resolution shell.

Data set	INS_L1_1	INS_L1_2	INS_L1_3	INS_L1_4	INS_L1_5	INS_L1_6	INS_L1_7	INS_L1_8	INS_L2_1	INS_L2_2	INS_L3_1	INS_L3_2	INS_cluster 1	INS_all
Unit-cell parameter a (Å)	78.0	77.9	78.0	78.0	78.0	78.0	77.9	77.9	78.5	78.9	78.2	78.3	78.0	78.0
Multiplicity	6.1 (6.1)	6.1 (6.0)	6.0 (6.0)	6.2 (6.2)	6.1 (6.2)	6.1 (6.1)	6.1 (6.1)	6.1 (6.1)	6.2 (5.7)	6.1 (5.7)	6.1 (6.2)	6.1 (5.9)	48.7 (47.3)	72.8 (67.2)
Completeness (%)	99.6 (100.0)	99.4 (96.5)	99.9 (99.8)	99.9 (100.0)	99.6 (97.1)	99.2 (96.8)	99.9 (100.0)	99.9 (100.0)	100.0 (100.0)	99.8 (97.4)	99.9 (100.0)	99.7 (100.0)	100.0 (100.0)	100.0 (100.0)
R factor† (%)	4.2 (12.9)	4.0 (12.1)	4.0 (11.0)	4.9 (16.6)	4.1 (12.5)	4.8 (15.6)	3.5 (9.1)	3.9 (11.5)	8.2 (51.8)	11.4 (62.1)	4.0 (10.5)	4.2 (11.2)	5.5 (14.3)	13.9 (26.7)
$\langle I/\sigma(I)\rangle$	32.3 (13.2)	33.7 (13.6)	34.9 (14.8)	28.3 (10.5)	33.4 (13.7)	29.7 (11.3)	38.1 (17.1)	34.1 (14.5)	22.5 (3.2)	16.5 (3.1)	33.3 (15.4)	35.3 (17.9)	74.0 (32.0)	52.1 (24.1)
B factor (Å^2)	23.9	23.9	23.8	24.4	23.3	24.1	23.5	23.6	27.1	26.1	24.2	24.4	23.8	24.1
$\langle \Delta F/\sigma(\Delta F)\rangle$, first shell	1.9	2.2	2.1	1.8	2.3	1.9	2.3	1.9	1.6	1.3	1.7	1.7	3.9	2.8
CC$_{\rm ano}$, first shell (%)	71	87	75	76	89	84	90	83	75	76	86	88	94	52
Best CC$_{\rm all}$ (%)	28.2	32.9	29.7	28.0	41.9	29.9	45.9	37.0	31.5	31.2	38.2	38.3	47.7	48.8
Best CC$_{\rm weak}$ (%)	13.8	8.6	14.2	18.7	17.8	13.4	29.1	15.8	14	10.2	15.1	17.4	26.5	24.1
No. of solutions	0	0	0	0	28	0	53	2	0	0	2	2	363	80
CC$_{\rm map}$ (%)	0	0	0	0	84	0	84	80	0	0	80	81	86	84

† R factor = $\sum[|I(h,i) - I(h)|]/\sum I(h,i)$ calculated using *XSCALE*.

containing every element at the other. Clustering thus starts with each element as an effectively separate cluster and then merges them into successively large clusters.

To perform the analysis, the distance between the clusters has to be defined and calculated. Here, when applied to diffraction data obtained from frozen crystals, the clustering technique is applied through use of the correlation coefficient (CC) matrix. This matrix is generated according the CC$_I(i, j)$ values output by *XSCALE* (Kabsh, 2010*b*) calculated in resolution shells for the common unique intensities of each pair of data sets. The mean intensity in each shell for each data set is subtracted (W. Kabsh, private communication). The CC$_I(i, j)$ values are converted to a distance matrix using the equation

$$d(i,j) = [1 - \mathrm{CC}_I^2(i,j)]^{1/2}. \qquad (1)$$

The distance $L(A, B)$ between two generic clusters A and B is defined using the average linkage. This is given by the formula

$$L(A,B) = \frac{1}{N_A N_B}\sum_{p=1}^{N_A}\sum_{q=1}^{N_B} d(i_p, j_q), \qquad (2)$$

where N_A and N_B are the number of elements in each cluster and i_p and j_q are the data sets in each of them.

To select the number of clusters, it is possible to cut the dendrogram at a specific level of similarity. We applied the identical cluster cutoff distance L = 0.14 to all four systems studied. This corresponds to a correlation coefficient CC$_I(i, j)$ of 0.99. The data sets were produced by scaling (using *XSCALE*) and merging together the data forming the largest cluster. A further data set for each test system was produced by scaling all data together. It may be possible that application of more sophisticated clustering methods would allow automation of the distance-cutoff criterion; however, use of the correlation coefficient seems a sufficiently intuitive method for this initial work.

The practical implementation of this cluster analysis was performed using the R language for statistical computing (R Development Core Team, 2011).

2.4. Substructure determination and phasing

For experimental phasing, the *SHELXC*/*D*/*E* program suite was used through the *HKL2MAP* interface (Pape & Schneider, 2004; Sheldrick, 2008). The signal-to-noise ratio in anomalous differences, $\langle \Delta F/\sigma(\Delta F)\rangle$ and the correlation coefficient between the anomalous differences in a randomly split data set CC$_{\rm ano}$ (Schneider & Sheldrick, 2002), as output by the program *SHELXC*, were used as initial indicators of data quality. On the basis of these statistics, the resolution cutoffs recommended by Schneider & Sheldrick (2002) were applied for substructure solution with *SHELXD*. In all cases presented below the results of substructure solution and phasing, whether successful or unsuccessful, were stable with respect to the resolution cutoff.

For all data sets, 1000 trials of substructure solution were used in *SHELXD* to find the anomalous scatterers. These were six sulfur sites for insulin, six disulfide superatoms and one calcium site for trypsin, 17 sulfur sites for thaumatin and 11 selenium sites for AroF. The success rate for each case was determined using a selected CC_{all}/CC_{weak} cutoff of 35/15 for insulin, 38/12 for trypsin, 40/15 for thaumatin and 35/15 for AroF. The cutoffs were selected by analyzing the CC_{all} distribution of the best data set for each of the four proteins. The standard *SHELXD* statistics CC_{all}/CC_{weak} were used to evaluate the data-set performance in the substructure solution (Sheldrick, 2010; Schneider & Sheldrick, 2002). In order to assess the accuracy of the final experimental phases, we used the map correlation coefficients (CC_{map}) between the electron-density maps calculated using the phases output by *SHELXE* and the phases calculated from the model. The PDB entries used were 2bn3 and 2bnl (Nanao *et al.*, 2005) for insulin and thaumatin, respectively, 2g55 (Mueller-Dieckmann *et al.*, 2007) for trypsin and 2o11 (M. Bruning, G. P. Bourenkov, N. I. Strizhov & H. D. Bartunik, unpublished work) for AroF. The model phases were estimated using the program *REFMAC*5 (Murshudov *et al.*, 2011) after appropriate origin matching between the model and the *SHELXD*/*E* solution, but without any refinement applied to the model. The procedure is justified by the fact that the PDB models used were refined to much higher resolution compared with the resolution of our data and there were no significant changes in the CC_{map} after refinement was carried out for a few selected data sets.

3. Results

3.1. Insulin

The small protein insulin is often used to test the ability to solve MX structures by sulfur SAD. For this study, we selected really tiny crystals and an X-ray energy that was far from the optimal energy usually recommended for sulfur SAD measurements (Mueller-Dieckmann *et al.*, 2005). Based on common standards, such as resolution, merging statistics and completeness, all 12 diffraction data sets collected are of rather good quality, with some variations in the statistical parameters owing to different crystal sizes and diffraction quality (Table 2). The data-collection redundancy was relatively low for a crystal with cubic symmetry and the anomalous difference signal was weak even in the low-resolution shells. Nevertheless, a clear substructure solution was obtained using some of the individual data sets, although at a fairly low

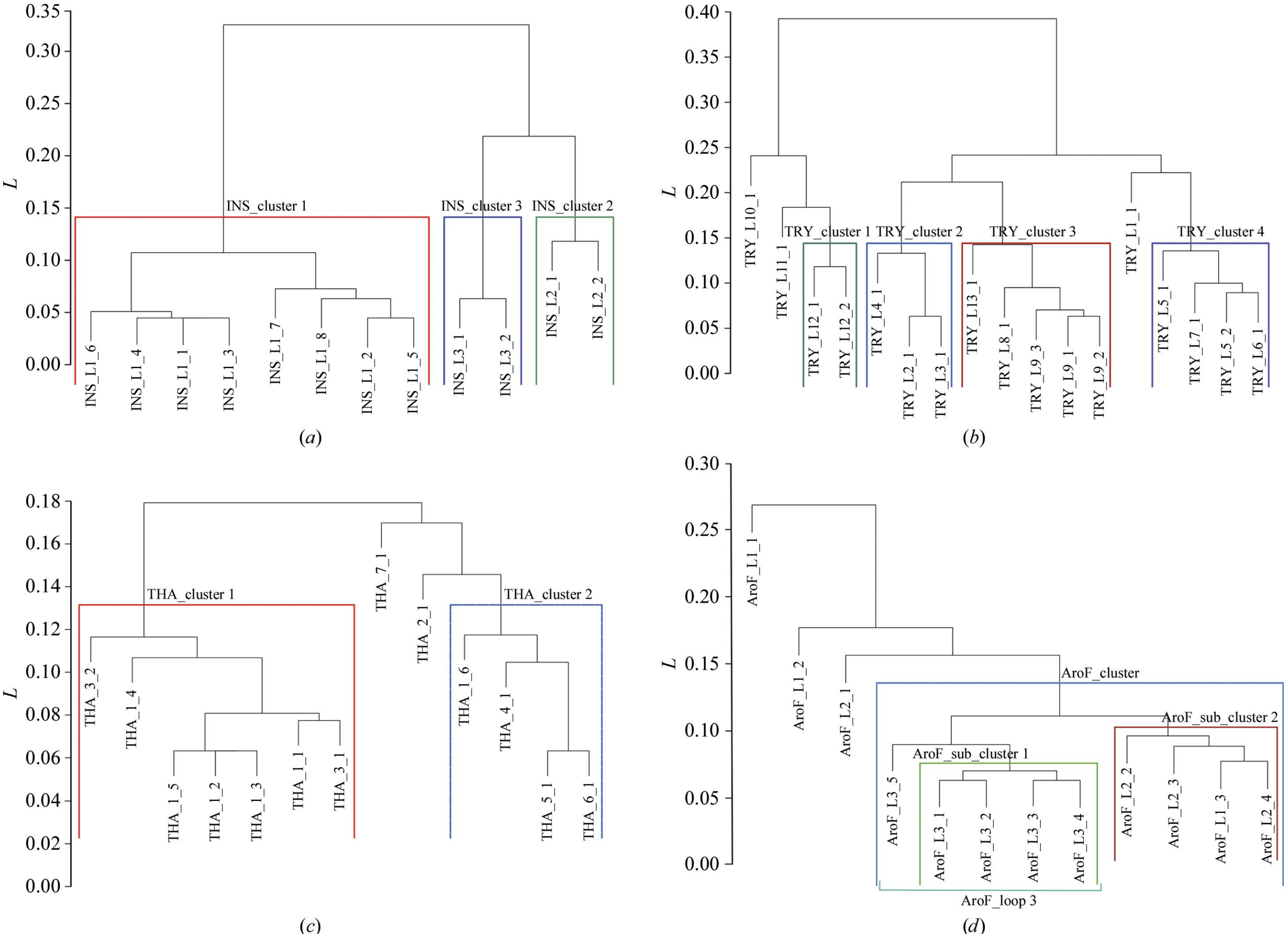

Figure 1
Hierarchical cluster-analysis dendrograms for the insulin (*a*), trypsin (*b*), thaumatin (*c*) and AroF (*d*) data sets.

Table 3
Trypsin: statistics for individual data sets, for the cluster and for all data sets merged.

Values in parentheses are for the highest resolution shell.

Data set	TRY_L1_1	TRY_L2_1	TRY_L3_1	TRY_L4_1	TRY_L5_1	TRY_L5_2	TRY_L6_1	TRY_L7_1	TRY_L8_1	TRY_L9_1	TRY_L9_2	TRY_L9_3	TRY_L10_1	TRY_L11_1	TRY_L12_1	TRY_L12_2	TRY_L13_1	TRY_cluster 3	TRY_all
Unit-cell parameters (Å)																			
a	54.65	54.73	54.63	54.63	54.62	54.63	54.71	54.51	54.72	54.48	54.50	54.68	54.50	54.44	54.55	54.55	54.41	54.72	54.65
c	107.31	107.61	107.50	107.46	107.46	107.43	107.58	107.00	107.82	107.38	107.48	107.81	106.66	107.09	107.19	107.28	107.46	107.82	107.31
Multiplicity	10.7 (10.1)	11.2 (11.4)	11.2 (11.4)	11.3 (11.4)	11.0 (10.4)	11.0 (10.6)	10.9 (10.5)	11.0 (10.8)	10.8 (10.5)	11.4 (11.6)	11.3 (11.3)	11.1 (10.8)	11.2 (11.1)	11.0 (10.5)	10.5 (9.9)	10.8 (10.3)	11.2 (11.0)	55.3 (52.3)	186.1 (178.0)
Completeness (%)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	98.9 (98.2)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)
R factor (%)	6.6 (15.0)	4.9 (8.4)	5.6 (9.1)	6.4 (13.1)	6.1 (14.9)	5.8 (11.5)	5.8 (11.4)	6.1 (10.5)	5.9 (10.8)	5.8 (12.0)	6.7 (16.4)	8.5 (23.4)	10.3 (26.5)	5.1 (10.7)	9.9 (25.4)	5.4 (12.9)	6.5 (12.3)	8.6 (18.0)	16.8 (26.7)
$\langle I/\sigma(I)\rangle$	29.0 (14.3)	39.0 (23.9)	34.5 (21.7)	31.2 (16.9)	32.9 (15.6)	33.2 (17.8)	32.8 (18.0)	31.3 (18.4)	32.1 (18.8)	33.7 (17.5)	30.4 (13.9)	24.4 (10.5)	22.4 (10.2)	38.1 (20.2)	21.6 (9.9)	35.8 (17.3)	30.5 (16.9)	53.4 (27.8)	57.6 (33.7)
B factor (Å^2)	22.0	21.7	21.2	22.7	22.4	21.2	21.5	20.4	21.2	21.9	22.1	23.4	23.4	20.2	23.1	21.2	20.3	21.5	21.1
$\langle \Delta F/\sigma(\Delta F)\rangle$, first shell	1.2	1.5	1.3	1.2	1.4	1.4	1.4	1.4	1.3	1.4	1.4	1.1	1.2	1.6	0.9	1.4	1.1	2.2	2.2
CC$_{\rm ano}$, first shell (%)	73	68	70	60	82	80	87	90	79	75	64	65	68	76	62	75	70	85	87
Best CC$_{\rm all}$ (%)	30.2	29.3	30.5	29.5	33.1	31.7	33.9	32.4	31.1	32.4	29.2	30.1	30.4	30.2	28.2	30.3	29.5	45.2	51.7
Best CC$_{\rm weak}$ (%)	12.5	10.5	6.4	9.7	12.7	3.6	10.4	6.5	4.2	6.8	7.3	12.2	9.9	7.9	11.3	6.8	8.2	19.5	25.9
No. of solutions	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	37
CC$_{\rm map}$ (%)	0	0	3	0	0	0	1	3	0	2	0	1	2	0	0	0	0	29	34

contrast level, as can be judged by both the low frequency of hits (0.1–0.5%) and the low CC$_{\rm weak}$ scores (<20%) of the top solutions. Only one data set, INS_L1_7, provided a convincing solution on the CC$_{\rm all}$/CC$_{\rm weak}$ criterion (Schneider & Sheldrick, 2002). Apart from this data set, which had a significantly lower overall $R_{\rm merge}$ of 0.035, the data sets delivering solutions could not be distinguished on the basis of the standard data statistics.

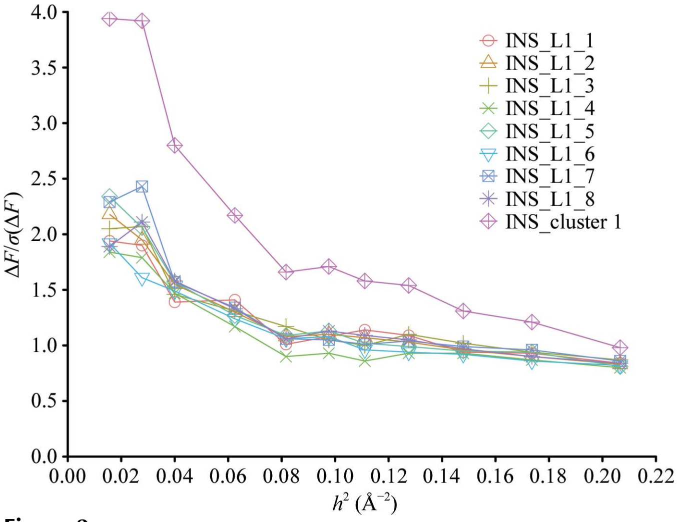

Figure 2
Anomalous difference ratio as a function of the squared scattering vector h^2 [h^2 = $|S^2|$ = $(2\sin\theta/\lambda)^2$] for individual data sets and for the cluster identified. The figure was produced using *ggplot2* (Wickham, 2009).

The results of the hierarchical cluster analysis are shown in Fig. 1(*a*). From this dendrogram, it is possible to distinguish three principal clusters, named INS_clusters 1, 2 and 3. The most prominent result of the analysis is that all of the crystals mounted in the same loop appear in the same cluster. Eight data sets constituting the largest cluster were merged together to produce the data set INS_cluster 1.

The anomalous signal statistics $\langle \Delta F/\sigma(\Delta F)\rangle$ of the merged data were improved considerably with respect to any of the individual data sets (Fig. 2). The improvement is most evident as an increase in the $\langle \Delta F/\sigma(\Delta F)\rangle$ and CC$_{\rm ano}$ statistics in the lowest resolution shells (Table 2). When N ideally isomorphous data sets with approximately equal $\langle \Delta F/\sigma(\Delta F)\rangle$ are merged, the resulting anomalous signal $\langle |\Delta F|/\sigma(\Delta F)\rangle_N$ can be estimated according to the approximate equation

$$N = \frac{\left\langle \dfrac{\Delta F}{\sigma(\Delta F)} \right\rangle_N^2 - \dfrac{2}{\pi}}{\left\langle \dfrac{\Delta F}{\sigma(\Delta F)} \right\rangle^2 - \dfrac{2}{\pi}}. \qquad (3)$$

We derive (3) under the assumption that ΔF is given by the convolution of anomalous difference and normally distributed random error. The estimate is only applicable when signals are significant: $\langle \Delta F/\sigma(\Delta F)\rangle^2 > 2/\pi$. For INS_cluster 1 with eight

Table 4
Thaumatin: statistics for individual data sets, for the cluster and for all data sets merged.

Values in parentheses are for the highest resolution shell.

Data set	THA_L1_1	THA_L1_2	THA_L1_3	THA_L1_4	THA_L1_5	THA_L1_6	THA_L2_1	THA_L3_1
Unit-cell parameters (Å)								
a	58.00	57.92	57.92	57.92	57.88	57.93	57.95	57.93
c	150.35	150.23	150.22	150.28	150.22	150.32	150.27	150.23
Multiplicity	8.0 (7.7)	8.0 (7.5)	8.0 (7.6)	8.0 (7.7)	8.0 (7.6)	8.0 (7.6)	8.0 (7.6)	8.0 (7.5)
Completeness (%)	99.9 (98.1)	99.9 (99.3)	99.9 (98.2)	99.9 (99.3)	99.9 (98.1)	99.9 (98.8)	100.0 (100.0)	99.9 (99.6)
R factor (%)	6.9 (23.9)	8.2 (30.0)	10.0 (33.8)	7.9 (28.5)	6.9 (23.9)	9.2 (37.5)	6.6 (22.9)	6.4 (20.5)
$\langle I/\sigma(I)\rangle$	26.8 (8.5)	23.9 (6.5)	21.9 (6.6)	25.5 (7.8)	26.8 (8.5)	21.8 (5.7)	27.9 (8.4)	27.8 (9.4)
B factor (Å^2)	13.3	13.6	13.6	13.8	13.3	14.0	13.5	13.0
$\langle \Delta F/\sigma(\Delta F)\rangle$, first shell	2.2	2.2	2.0	2.3	2.4	2.1	2.5	2.4
CC$_{\rm ano}$, first shell (%)	76	75	79	81	87	79	84	82
Best CC$_{\rm all}$ (%)	43.8	50.3	41.5	45.5	45.3	43.3	46.4	47.3
Best CC$_{\rm weak}$ (%)	19.9	20.6	16.2	21.6	20.0	18.7	22.4	23.8
No. of solutions	26	38	1	38	44	16	17	58
CC$_{\rm map}$ (%)	0	15	6	24	20	35	33	46

Data set	THA_L3_2	THA_L4_1	THA_L5_1	THA_L6_1	THA_L7_1	THA_cluster 1	THA_all
Unit-cell parameters (Å)							
a	57.98	57.84	57.81	57.82	57.89	58.00	58.00
c	150.12	150.08	150.10	150.17	150.21	150.35	150.35
Multiplicity	8.3 (7.6)	6.4 (1.3)	8.0 (7.7)	8.0 (7.9)	8.0 (7.7)	56.0 (54.4)	100.9 (94.7)
Completeness (%)	96.3 (96.4)	90.0 (22.4)	100.0 (99.9)	100.0 (100.0)	100.0 (100.0)	99.9 (99.9)	100.0 (100.0)
R factor (%)	8.9 (32.6)	4.9 (17.1)	4.7 (11.8)	7.3 (13.5)	15.7 (46.7)	9.2 (18.7)	12.2 (23.6)
$\langle I/\sigma(I)\rangle$	22.2 (6.4)	27.7 (2.6)	35.0 (14.8)	23.8 (13.8)	15.0 (4.5)	59.2 (30.9)	57.3 (31.9)
B factor (Å^2)	13.0	14.4	12.6	13.0	15.4	12.9	12.8
$\langle \Delta F/\sigma(\Delta F)\rangle$, first shell	2.1	2.6	2.7	2.1	1.9	4.5	3.9
CC$_{\rm ano}$, first shell (%)	76	78	87	81	65	94	96.4
Best CC$_{\rm all}$ (%)	32.3	44.7	50.1	35.5	34.9	57.6	55.1
Best CC$_{\rm weak}$ (%)	12.6	24.1	22.3	13.3	8.9	32.3	26.6
No. of solutions	0	20	143	0	0	323	181
CC$_{\rm map}$ (%)	2	30	80	9	1	83	60

data sets, the increase in $\langle \Delta F/\sigma(\Delta F)\rangle$ is close to that expected for an ideal case. In the subsequent analysis, in order to characterize the signal to noise in anomalous differences we used $\langle \Delta F/\sigma(\Delta F)\rangle$ and CC$_{\rm ano}$ statistics for reflections in the first resolution shell >8 Å (Tables 2, 3 and 4).

The striking improvement in the data is clearly reflected by the success rate of the substructure solution with *SHELXD*. For INS_cluster 1 data a high CC$_{\rm all}$/CC$_{\rm weak}$ scored solution was obtained approximately every third trial. Adding the four residual data sets to INS_cluster 1 and producing the INS_all merged data set in fact significantly reduced the anomalous signal in the low-resolution shell, in particular with respect to the CC$_{\rm ano}$ parameter. The substructure solution was still successful, although the success rate dropped considerably compared with the INS_cluster 1.

Remarkably, for insulin all of the successful substructure solutions in *SHELXD* resulted in highly accurate sets of phases output by *SHELXE* as indicated by CC$_{\rm map.}$

3.2. Trypsin

The anomalous signal for trypsin crystals is weaker compared with the signal from insulin crystals (Table 1). Data-processing statistics for TRY data sets are presented in Table 3. The diffraction quality varied significantly from crystal to crystal [$0.5 < R_{\rm merge} < 0.10$, $8 <$ last shell $I/\sigma(I) < 20$]. None of the individual data sets had a significant level of $\langle \Delta F/\sigma(\Delta F)\rangle$, even at low resolution. Consequently, none of the substructure solution attempts with individual data sets was successful.

The cluster dendrogram of TRY data (Fig. 1*b*) revealed appreciable non-isomorphism, as indicated by the highest overall cluster distance $L = 0.4$ among the four systems studied. Four rather small principal clusters having from two to five data sets were separated at the threshold level $L = 0.14$. The statistics of the merged data set in the largest cluster TRY_cluster 3 improved significantly; the increase in the low-resolution $\langle \Delta F/\sigma(\Delta F)\rangle$ value is consistent with that expected for the fivefold increase in the multiplicity. For the substructure solution we exploited the data truncated at 3.5 Å resolution and searched for six disulfides as superatoms and one calcium site (the *SHELXD* settings were identical to those used for individual TRY data sets). The solution was unambiguously successful according to the CC$_{\rm all}$/CC$_{\rm weak}$ criterion, although the success rate was still rather low. The resulting electron-density maps were rather noisy overall, as expected for very weak SAD data and low solvent content (32%), but would nevertheless provide a starting point for successful model building according to a visual inspection of the maps.

Despite the large total number of TRY data sets, merging all of them together did not result in any significant

Table 5
AroF statistics for individual data sets, for the cluster and for all data sets merged.

Values in parentheses are for the highest resolution shell.

Data set	AroF_L1_1	AroF_L1_2	AroF_L1_3	AroF_L2_1	AroF_L2_2	AroF_L2_3	AroF_L2_4
Unit-cell parameters (Å)							
a	132.38	132.15	132.93	132.03	132.98	132.03	132.07
c	161.01	160.57	160.25	160.26	160.10	160.32	160.46
Multiplicity	14.0 (13.1)	11.6 (11.4)	11.6 (11.5)	11.5 (11.2)	11.5 (11.2)	11.5 (11.4)	11.5 (11.3)
Completeness (%)	82.6 (84.6)	99.2 (99.6)	99.9 (100.0)	100.0 (100.0)	99.9 (100.0)	99.9 (100.0)	100.0 (100.0)
R factor (%)	33.5 (87.7)	27.9 (81.2)	11.3 (26.6)	23.3 (66.7)	13.6 (32.5)	12.6 (28.4)	14.6 (31.6)
$\langle I/\sigma(I)\rangle$	8.5 (3.3)	10.3 (4.1)	20.8 (10.0)	11.5 (4.54)	17.1 (8.8)	17.5 (9.3)	14.8 (7.9)
B factor (Å^2)	32.6	36.6	29.3	34.2	31.0	30.6	30.3
$\langle \Delta F/\sigma(\Delta F)\rangle$, first shell	2.4	2.7	4.9	2.9	3.6	3.4	3.1
CC$_{\rm ano}$, first shell (%)	77	84	96	90	95	94	94
Best CC$_{\rm all}$ (%)	20.9	42.8	56.5	44.1	52.7	50.2	50.3
Best CC$_{\rm weak}$ (%)	7.8	25.9	28.8	23.5	26.3	23.2	25.7
No. of solutions	0	29	372	51	201	137	147
CC$_{\rm map}$ (%)	59	80	89	86	88	88	87

Data set	AroF_L3_1	AroF_L3_2	AroF_L3_3	AroF_L3_4	AroF_L3_5	AroF_cluster	AroF_all
Unit-cell parameters (Å)							
a	132.15	132.08	132.17	132.16	132.21	132.93	132.38
c	160.71	160.53	160.77	160.78	160.83	160.25	161.01
Multiplicity	11.6 (11.5)	11.6 (11.5)	11.5 (11.5)	11.5 (11.5)	11.5 (11.5)	103.8 (103.8)	137.5 (125.7)
Completeness (%)	99.6 (99.9)	99.9 (100.0)	100.0 (100.0)	99.9 (100.0)	100.0 (100.0)	100.0 (100.0)	99.9 (99.9)
R factor (%)	12.6 (30.6)	10.8 (23.7)	8.5 (15.9)	9.0 (17.3)	8.7 (16.2)	12.4 (26.0)	17.1 (38.4)
$\langle I/\sigma(I)\rangle$	21.7 (12.2)	21.1 (11.0)	24.9 (14.9)	24.1 (14.5)	24.7 (15.1)	58.9 (34.5)	60.7
B factor (Å^2)	29.9	29.7	29.1	28.1	28.9	28.7	32.6
$\langle \Delta F/\sigma(\Delta F)\rangle$, first shell	4.5	4.4	4.4	4.5	4.3	11.3	11.9
CC$_{\rm ano}$, first shell (%)	95	95	95	96	94	99	99
Best CC$_{\rm all}$ (%)	57.2	55.1	56.0	55.6	56.2	61.0	61.2
Best CC$_{\rm weak}$ (%)	35.3	27.0	27.7	27.6	29.2	32.6	32.2
No. of solutions	404	294	350	305	384	524	491
CC$_{\rm map}$ (%)	88	88	88	89	88	89	89

improvement in the anomalous data statistics. The contrast and success rate of structure solution, as well as the accuracy of the resulting phases, improved marginally (Table 3).

3.3. Thaumatin

For data collection from thaumatin crystals we used a lower X-ray energy and the expected anomalous signal should be stronger than in the insulin and trypsin cases (Table 1). In spite of the small size of the crystals, all data sets showed rather good statistical quality and were similar to each other in quality. An exception is THA_L5_1, where the crystal used diffracted significantly more strongly than the others. The substructure searches were unsuccessful for only four data sets (Table 4), but only THA_L5_1 also produced a high-quality electron-density map. Other maps had rather poor contrast, as indicated by low CC$_{\rm map}$ values. The results of hierarchical cluster analysis applied to the THA data sets are presented in Fig. 1(*c*).

Crystals of thaumatin appeared to be less divergent structurally compared with insulin or trypsin crystals. This is indicated by the mean cluster distance value of L = 0.18 for all data sets. The largest THA_cluster 1 contained seven data sets collected from crystals mounted in different loops: THA_L1 and THA_L3.

With the merged THA_cluster 1 data set the anomalous signal increased quantitatively as expected for the seven isomorphous data sets. The success rate of substructure solution rose above 30% and excellent-quality phase sets were obtained from *SHELXE*. Note that this cluster did not include the best phasing data set THA_L5_1. With all THA data sets merged together, the success rate and contrast in the substructure solution decreased, although the solution was still unambiguous. The quality of the resulting electron-density maps was clearly compromised compared with either the best cluster or the best individual data set.

3.4. AroF

AroF crystals diffracted to a much lower resolution than the crystals of all other systems in this study; however, the Se atoms provide a much stronger anomalous signal compared with sulfur (Tables 1 and 5). Table 5 presents the data-processing statistics for single AroF data sets. There are variations in data quality that are a consequence of different crystal sizes and diffraction quality. Although the regular data statistics [$R_{\rm merge}$ and $I/\sigma(I)$] for some of the data sets appear at the limit of being acceptable (*e.g.* overall $R_{\rm merge}$ > 0.3), the $\langle \Delta F/\sigma(\Delta F)\rangle$ was rather high and the Se-substructure determination was successful for all individual data sets.

Hierarchical cluster analysis of the AroF data is presented in Fig. 1(*d*). Most of the crystals are isomorphous. A large cluster consisting of nine (out of 12) data sets can be selected for the mean cluster distance cutoff L = 0.14. Merging together the data set of the AroF_cluster improved the signal to noise in anomalous differences nearly threefold to the unusually high level of >10. This improvement was clearly reflected by the substructure-solution statistics (a success rate of >50% and high CC_{all}/CC_{weak} values), but an improvement in terms of electron-density maps, which already had very high quality for most of the individual AroF data sets, was not visible. Using all data, *i.e.* adding three data sets lying outside the principal cluster, did not have any qualitative influence on the data.

4. Discussion

The effects of radiation damage in macromolecular crystallography cannot be overcome. We can optimize the conditions of measurement, but there is an absolute limit to the resolution and data quality which can be obtained from one crystal (Bourenkov & Popov, 2010). The situation can only be improved by merging data from a subset of homogeneous samples. This approach can be applied to assemble a complete data set and to increase the signal-to-noise level of experimental data. As has recently been demonstrated experimentally (Liu *et al.*, 2011), the anomalous signal, success in SAD substructure determination and accuracy of phases and electron-density maps all improve with an increase in the number of crystals used in merging. The results of our tests completely support these conclusions, with the condition that merging of isomorphous data sets has taken place.

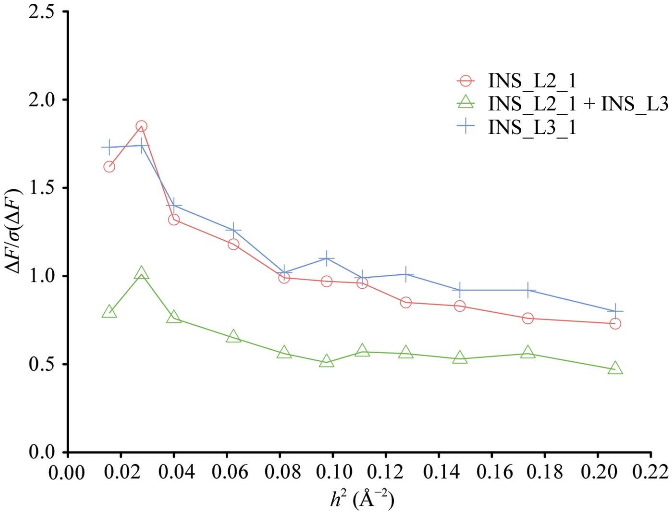

Figure 3
Anomalous difference ratio as a function of the squared scattering vector h^2 for data sets INS_L2_1 and INS_L3_1 and for the merged data. Merging data from crystals identified as belonging to separate clusters results in a reduction of the anomalous signal (Wickham, 2009).

In this study, we evaluated the possibility of using hierarchical cluster analysis using the cluster-distance metric based on the intensity correlation coefficients as a tool for such a selection. The results clearly demonstrate improvement in the quality of the multi-crystal anomalous difference data sets created on the basis of cluster analysis in all of our tests. How strong the effects of the improvement were on the different steps of the structure solution varied significantly from case to case. This is not surprising in view of the rather high complexity of the phasing process, which is dependent on many parameters in addition to the signal-to-noise ratio in the data. As a further level of complexity, the comparative results presented here are influenced by the particular sample of structural variability that was captured in each experiment. Nevertheless, we demonstrate that in none of the examples were the best cluster-based multi-crystal data of inferior quality when compared with any of the single-crystal data sets or compared with all data sets merged blindly with no analysis.

The addition of non-isomorphous data sets significantly deteriorated (in the cases of insulin and thaumatin) or did not further improve (for AroF) the anomalous signal. In the case of trypsin, where only a small isomorphous cluster was found, averaging all data gave slightly better results. The necessity of selecting isomorphous subsets when using the multi-crystal data is best illustrated using the examples involving a small number of data sets. Fig. 3 presents the $\langle \Delta F/\sigma(\Delta F)\rangle$ values plotted against the resolution shells for two data sets INS_L2_1 and INS_L3_1 that fall into different clusters and for a data set where the two are merged together. In the latter, the weak anomalous signal which was initially present deteriorates completely. Fig. 4 represents the electron-density map

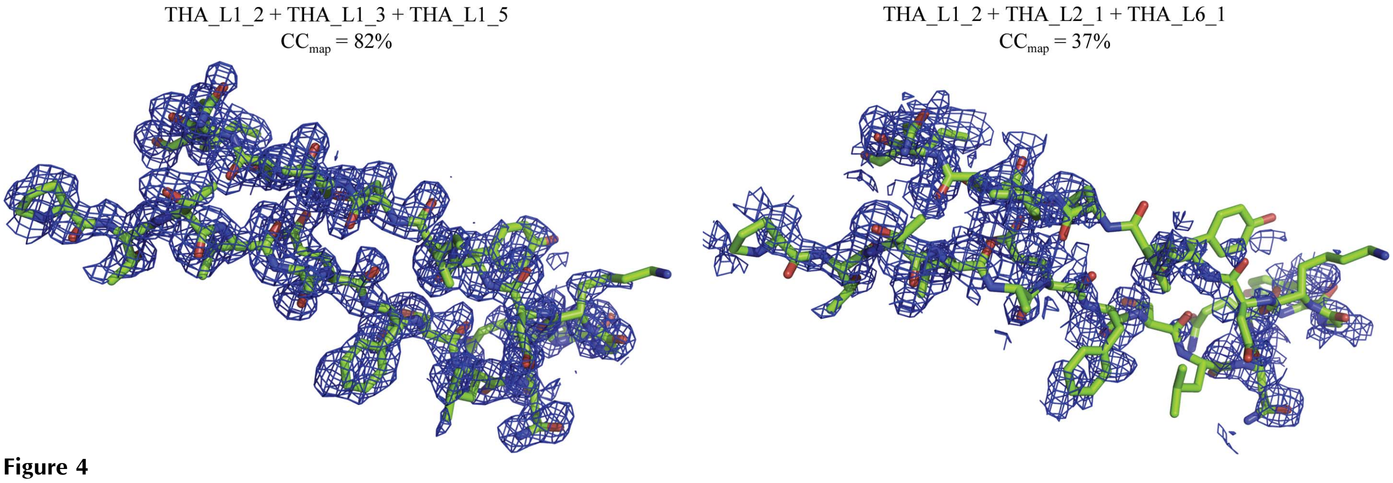
THA_L1_2 + THA_L1_3 + THA_L1_5
CC_{map} = 82%
THA_L1_2 + THA_L2_1 + THA_L6_1
CC_{map} = 37%

Figure 4
Electron-density maps for thaumatin after *SHELXE* obtained using combined data sets.

for thaumatin data sets calculated using *SHELXE* phases after three cycles of autotracing. The map obtained by merging data sets belonging to the same cluster, THA_L1_2, THA_L1_3 and THA_L1_5, presents a marked improvement in the phases, with a CC_{map} of 82%. In contrast, after merging data sets THA_L1_2, THA_L2_1 and THA_L6_1 the map remains uninterpretable, with a CC_{map} of 37%. In the following, we address the question of whether our analysis does in fact deliver optimal data subsets or whether a better selection would be possible. Being unable to carry out an exhaustive analysis of all possible subsets of data, we followed the success rate of the thaumatin substructure solution as a function of the number of data sets included in merging, starting with THA_L1_2 and first adding the data sets inside the principal THA_cluster 1 and then expanding the cluster. The result is plotted in Fig. 5. One can clearly see that the data quality improves continuously while adding more data sets that belong to the cluster. Further addition of the (distinctly best) data set still improves the contrast (marginally), but adding further data sets degrades the results continuously. Similar analysis for the AroF data is presented in Fig. 6. Here, contrast saturation is rapidly achieved inside the small AroF_sub_clusters 1 and 2, each with four members and separated at the threshold $L = 0.10$. The contrast remained practically unchanged on adding the ninth member of the principal cluster and then marginally degraded on adding further data outside the cluster.

Our results support the choice of the correlation coefficient between data sets and the cluster distance described in (1) as a reliable indicator of crystal isomorphism. The choice of the clustering threshold value was based on experience and was also proven to be practically useful. Both the choice of the metrics and the best cluster-selection procedure can be optimized for various problems and conditions. The advantages of the intensity-based correlation metric chosen are: (i) it directly uses the standard data-processing program (*XSCALE*) output, (ii) it is effectively independent of the scaling the data sets together and on the estimation of standard uncertainties and (iii) the ready application of the same method to the study of isomorphism beyond anomalous scattering. One obvious drawback of the intensity correlation-based metric is that it does not account for the signal to noise in the data and may be less indicative when data quality varies strongly between the data sets. This is observed in case of AroF, where three data sets AroF_L1_1, AroF_L1_2 and AroF_L2_1 lying outside the principal cluster are much weaker than the others; when included in merging they are correctly down-weighted and induce no negative influence on the result. Improved metrics accounting for such effects can be derived.

For the SAD applications considered here, a metric based on the correlation between the signed ΔF values is clearly of interest. For instance, it would not only account for non-isomorphism and data errors, but also for the variation in the anomalous substructure in the case of heavy-atom derivatives. Unfortunately, such a simple statistic is not calculated by any of the standard scaling programs that we are aware of. One should also note that (1) is not applicable to low correlations on ΔF values, as expected for weak data. However, if in the future the appropriate analysis were made routinely available then the method described here would be routinely applicable.

The choice of clustering threshold criteria dependent on the expected signal could also improve the performance of the method in general. In fact, considering that the final aim of the clustering is to produce merged 'centroid' data, centroid-based clustering algorithms (*e.g.* Tan *et al.*, 2006) which do not require a notion of 'distance threshold' could be more efficient compared with the hierarchical clustering. Finally, a program selecting the best subset of the basis of global maximization of, for example, $\langle \Delta F/\sigma(\Delta F) \rangle$ on all possible data subsets is also feasible.

The experiments presented here, and the cluster analyses shown in Fig. 1, suggest a further rather important implication. In the cases of all four test systems, the principal clusters selected only on the basis of diffraction data consistency, without any knowledge of sample history, showed a very clear trend to group crystals mounted in the same cryoloop. This is

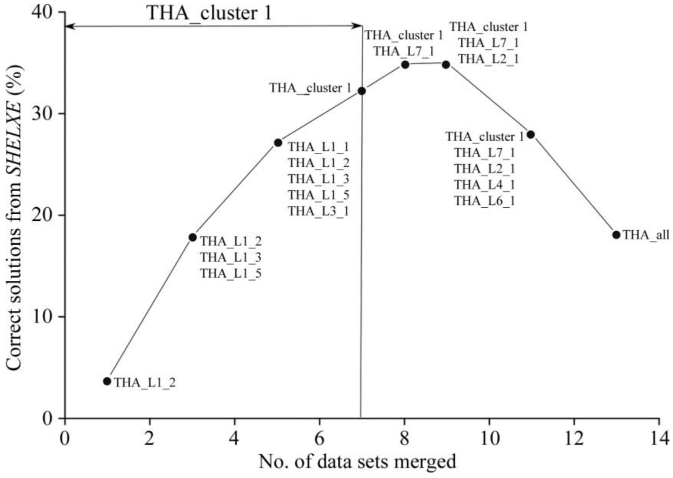

Figure 5
Percentage of successful solutions from *SHELXD versus* number of data sets merged for thaumatin.

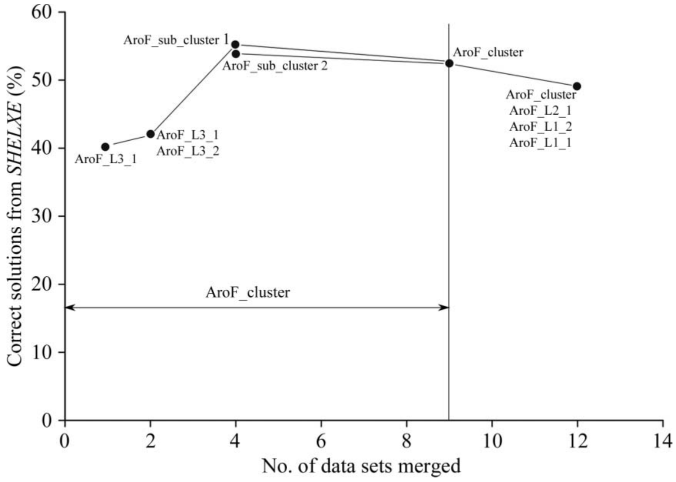

Figure 6
Percentage of successful solutions from *SHELXD versus* number of data sets merged for AroF.

precisely the case for all three clusters found for INS data. Crystals mounted on loop THA_L1 dominate in the principal cluster of thaumatin. The best phasing cluster of trypsin is mostly made up from crystals from loop TRY_L9. A similar situation is found for AroF_sub_clusters 1 and 2. These observations strongly suggest that the major physical source of the non-isomorphism is variation in sample treatment during sample transfer, cryoprotection, mounting and flash-cooling. Indeed, at least in our hands, sample treatment involved a number of poorly controlled parameters such as the soaking time in the cryosolution, the dehydration time during the mount transfer, the cooling rate in the cryostream *etc.* Interestingly, the overall best isomorphism, with the exception of the very weak data sets discussed above, is clearly observed for AroF crystals, which were treated without the cryoprotection step.

5. Conclusion

The purpose of the present study was to demonstrate the importance of hierarchical cluster analysis in understanding non-isomorphism of macromolecular crystals; in particular, it can help in the selection of isomorphous crystals. Determination of the distance matrix, based on the correlation coefficient between scaled intensities, played a major role in the present work. The results of this study illustrate that for SAD data collection it is of great importance to use a protocol that helps to appropriately merge the crystal data sets. The relevance of merging isomorphous data sets is clearly supported by the current findings for the four proteins investigated. A dendrogram was used to select the crystal data sets to improve the anomalous signal. The selection of data sets using this method was also advantageous in solving the crystal substructure for two of these proteins. Moreover, the improvement of the anomalous signal in SAD experiments obtained by merging isomorphous data could be advantageous, especially for weak signals.

For difficult crystallographic cases, this method could permit substructure determination for some cases where it is not possible using individual data sets. The results of this research support the idea that meticulous selection of data sets can lead to better determination of the substructure than using randomly chosen data sets.

The combination of multi-crystal data-collection techniques with advanced statistical data-analysis methods has clear potential to expand the applicability of sulfur-SAD phasing to more complex structures. This methodology should be further extended to aid other applications such as heavy-atom derivative phasing or, possibly, resolution enhancement for poorly ordered systems. The development of automated and reproducible sample-handling techniques that provide better control over the sample state throughout the process is likely to become another important component.

We thank Ulrike Kapp for provision of the test samples trypsin and thaumatin. Members of the ESRF Structural Biology Group provided constructive criticism and discussion during the development of this work. Portions of this work were supported by a European Framework Project grant (BIOXHIT; LSHG-CT-2003-503420).

References

Barr, G., Dong, W. & Gilmore, C. J. (2004). *J. Appl. Cryst.* **37**, 243–252.

Blundell, T. & Johnson, L. N. (1976). *Protein Crystallography.* New York: Academic Press.

Bourenkov, G. P. & Popov, A. N. (2010). *Acta Cryst.* D**66**, 409–419.

Buehler, A., Urzhumtseva, L., Lunin, V. Y. & Urzhumtsev, A. (2009). *Acta Cryst.* D**65**, 644–650.

Cianci, M., Helliwell, J. R. & Suzuki, A. (2008). *Acta Cryst.* D**64**, 1196–1209.

Dunlop, K. V., Irvin, R. T. & Hazes, B. (2005). *Acta Cryst.* D**61**, 80–87.

Garman, E. F. (2010). *Acta Cryst.* D**66**, 339–351.

Halle, B. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 4793–4798.

Hofmann, D. W. M., Kuleshova, L. N., Hofmann, F. & D'Aguanno, B. (2009). *Chem. Phys. Lett.* **475**, 149–155.

Juers, D. H. & Matthews, B. W. (2001). *J. Mol. Biol.* **311**, 851–862.

Kabsch, W. (1993). *J. Appl. Cryst.* **26**, 795–800.

Kabsch, W. (2010*a*). *Acta Cryst.* D**66**, 125–132.

Kabsch, W. (2010*b*). *Acta Cryst.* D**66**, 133–144.

Liu, Q., Zhang, Z. & Hendrickson, W. A. (2011). *Acta Cryst.* D**67**, 45–59.

Mueller-Dieckmann, C., Panjikar, S., Schmidt, A., Mueller, S., Kuper, J., Geerlof, A., Wilmanns, M., Singh, R. K., Tucker, P. A. & Weiss, M. S. (2007). *Acta Cryst.* D**63**, 366–380.

Mueller-Dieckmann, C., Panjikar, S., Tucker, P. A. & Weiss, M. S. (2005). *Acta Cryst.* D**61**, 1263–1272.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.

Nanao, M. H., Sheldrick, G. M. & Ravelli, R. B. G. (2005). *Acta Cryst.* D**61**, 1227–1237.

Nurizzo, D., Mairs, T., Guijarro, M., Rey, V., Meyer, J., Fajardo, P., Chavanne, J., Biasci, J.-C., McSweeney, S. & Mitchell, E. (2006). *J. Synchrotron Rad.* **13**, 227–238.

Paithankar, K. S., Owen, R. L. & Garman, E. F. (2009). *J. Synchrotron Rad.* **16**, 152–162.

Pape, T. & Schneider, T. R. (2004). *J. Appl. Cryst.* **37**, 843–844.

Ravelli, R. B. & Garman, E. F. (2006). *Curr. Opin. Struct. Biol.* **16**, 624–629.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. http://www.R-project.org.

Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.

Sheldrick, G. M. (2008). *Acta Cryst.* A**64**, 112–122.

Sheldrick, G. M. (2010). *Acta Cryst.* D**66**, 479–485.

Tan, P.-N, Steinbach, M. & Kumar, V. (2006). *Introduction to Data Mining.* Boston: Pearson Addison-Wesley.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer.