STRUCTURAL BIOLOGY

ISSN: 2059-7983
journals.iucr.org/d

Effect of impurities and post-experimental purification in SAD phasing with serial femtosecond crystallography data

Tao Zhang, Yuanxin Gu and Haifu Fan

Acta Cryst. (2016). D**72**, 789–794

STRUCTURAL BIOLOGY
ISSN 2059-7983

Effect of impurities and post-experimental purification in SAD phasing with serial femtosecond crystallography data

Tao Zhang, Yuanxin Gu and Haifu Fan*

Institute of Physics, Chinese Academy of Sciences, Beijing 100190, People's Republic of China. *Correspondence e-mail: fan@iphy.ac.cn

Received 7 December 2015
Accepted 16 April 2016

Edited by Q. Hao, University of Hong Kong

Keywords: serial crystallography; SFX; SAD phasing; effect of inpurities.

In serial crystallography (SX) with either an X-ray free-electron laser (XFEL) or synchrotron radiation as the light source, huge numbers of micrometre-sized crystals are used in diffraction data collection. For a SAD experiment using a derivative with introduced heavy atoms, it is difficult to completely exclude crystals of the native protein from the sample. In this paper, simulations were performed to study how the inclusion of native crystals in the derivative sample could affect the result of SAD phasing and how the post-experimental purification proposed by Zhang *et al.* [(2015), *Acta Cryst.* D**71**, 2513–2518] could be used to remove the impurities. A gadolinium derivative of lysozyme and the corresponding native protein were used in the test. Serial femtosecond crystallography (SFX) diffraction snapshots were generated by *CrystFEL*. *SHELXC*/*D*, *Phaser*, *DM*, *ARP*/*wARP* and *REFMAC* were used for automatic structure solution. It is shown that a small amount of impurities (snapshots from native crystals) in the set of derivative snapshots can strongly affect the SAD phasing results. On the other hand, post-experimental purification can efficiently remove the impurities, leading to results similar to those from a pure sample.

1. Introduction

Serial crystallography (SX) using either a hard X-ray free-electron laser (Chapman *et al.*, 2011) or third-generation synchrotrons (Gati *et al.*, 2014; Stellato *et al.*, 2014; Botha *et al.*, 2015; Nogly *et al.*, 2015) provides new opportunities in protein crystallography. It enables the solution of protein structures using multiple micrometre-sized crystals, while conventional protein crystal structure determination requires single crystals that are more than ten times larger. SX also greatly reduces the effect of radiation damage on the diffraction data and thus avoids the need to cryocool the sample. On the other hand, SX creates new problems. The use of a huge number of micrometre-sized crystals instead of only one or just a few single crystals leads to sample heterogeneity becoming a serious problem that affects the data quality and consequently the result of the analysis.

The SAD (single-wavelength anomalous diffraction) method, including the SAD phasing-based SIRAS (single isomorphous replacement with anomalous scattering) method (see Yao, Zhang *et al.*, 2014), is nowadays the first choice in solving protein structures that do not have appropriate known homologues. The first successful SFX SAD phasing was performed on an gadolinium derivative of hen egg-white lysozyme (Barends *et al.*, 2014). Yamashita *et al.* (2015) solved the structure of luciferin-regenerating enzyme (LRE) using SFX SIRAS data, but failed to solve the same structure *via* SFX SAD phasing. To the authors' knowledge, no previously

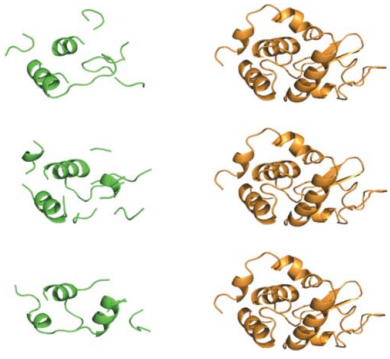

Table 1
Test samples.

The resolution range was 40–1.83 Å.

Sample No.	Impurities (%)	No. of snapshots	
		Derivative	Native
1	0	2.97×10^5	0
2	1	2.97×10^5	3000
3	2	2.97×10^5	6061
4	3	2.97×10^5	9185
5	4	2.97×10^5	12375
6	5	2.97×10^5	15631
7	6	2.97×10^5	18957
8	7	2.97×10^5	22354
9	8	2.97×10^5	25826
10	9	2.97×10^5	29373
11	10	2.97×10^5	33000
12	11	2.97×10^5	36707
13	12	2.97×10^5	40500
14	13	2.97×10^5	44379
15	14	2.97×10^5	48348
16	15	2.97×10^5	52411

unknown protein structures have been solved using SAD data from SX experiments. In addition to other reasons, this is owing to the fact that SAD signals are weak and sensitive to sample heterogeneity. In this paper, the effect of sample impurities on SFX SAD phasing is studied in detail using simulated data. In order to obtain sufficiently strong anomalous scattering signals to solve large protein structures by SAD phasing, derivative samples are often prepared by introducing heavy atoms into crystals of the native protein. In this case, there is no guarantee that the process is 100% complete. Hence, some native crystals would remain as 'impurities' in the derivative sample, which consists of a huge number of microcrystals. In the present test, we assumed that the samples are mainly derivative crystals but contain different percentages of native crystals as impurities. The method used to identify impurities and to purify the samples (Zhang *et al.*, 2015) is based on the variation of unit-cell parameters from native to derivative crystals. It is commonly found that the unit-cell parameters of native crystals are slightly different from those of derivative crystals [see, for example, Lu *et al.* (2011) and Yao, Cherney *et al.* (2014) with regard to single-crystal diffraction experiments and Botha *et al.* (2015) with regard to SX diffraction experiments]. In the present test, we assume that both the chemical and physical environments of the sample crystals are kept fixed during a single SFX experiment. Hence, the difference in the unit-cell parameters may be used to distinguish native crystals from derivative crystals.

2. Samples

The crystal structures of native hen egg-white lysozyme (PDB entry 4et8) and of a gadolinium derivative (PDB entry 4n5r) have been solved by Boutet *et al.* (2012) and Barends *et al.* (2014), respectively, using SFX data. The two structures were used in the calculation of simulated SFX diffraction snapshots. Hen egg-white lysozyme crystallizes in space group $P4_32_12$, with unit-cell parameters $a = 79.0$, $c = 38.0$ Å, and there are 129 amino-acid residues in the asymmetric unit. The Gd derivative of lysozyme crystallizes in space group $P4_32_12$, with unit-cell parameters $a = 79.1$, $c = 39.2$ Å, and there are 129 amino-acid residues and two Gd atoms in the asymmetric unit. SFX diffraction snapshots of the derivative and the native were calculated separately and then mixed in different percentages to prepare 16 sets of sample data as shown in Table 1. The simulation of SFX diffraction snapshots was performed using *CrystFEL* v.0.5.4a (White *et al.*, 2012) with the conditions summarized in Table 2.

Table 2
Conditions for the simulation of SFX diffraction snapshots.

Photoenergy (eV)	8.49×10^3
Photons per pulse	5×10^{12}
Pixel size (µm)	110
Pixels on the detector	1456 × 1456
Sample-to-detector distance (cm)	10
Crystal size (nm)	500–5000
Anomalous scattering factors of Gd	$f' = -0.996$, $f'' = 18.091$
Atomic parameters of the sample proteins	Taken from PDB entries 4n5r for the derivative and 4et8 for the native
Poisson noise (%)	6.9

Table 3
Numbers of snapshots of the native and the derivative included in the 'small' (blue) peak and the 'large' (red) peak for different samples after treatment by post-experimental purification.

Sample No.	Blue peak		Red peak	
	Native snapshots	Derivative snapshots	Native snapshots (% impurities)	Derivative snapshots
2	2960	1622	30 (0.01)	295124
3	5979	1622	62 (0.02)	295124
4	9074	1622	89 (0.03)	295124
5	12222	1622	122 (0.04)	295124
6	15449	1622	147 (0.05)	295124
7	18743	1622	173 (0.06)	295124
8	22106	1622	197 (0.07)	295124
9	25530	1622	239 (0.08)	295124
10	29034	1622	276 (0.09)	295124
11	32613	1622	315 (0.11)	295124
12	36277	1622	350 (0.12)	295124
13	40022	1622	390 (0.13)	295124
14	43838	1622	443 (0.15)	295124
15	47759	1622	484 (0.16)	295124
16	55891	1622	525 (0.18)	295124

3. Purification and diffraction-intensity extraction

Sample Nos. 2–16 in Table 3 were subjected to post-experimental purification. A scan of the primitive unit-cell volume was implemented for each sample. Fig. 1 shows the results for sample Nos. 3, 6, 9, 12 and 15. As can be seen, there are two peaks in each plot, indicating that each sample consists of two components. Their unit cells correspond to those of the Gd derivative (the red peak) and the native (the blue peak). A window is used to define the full width of each peak for each sample. The purification simply involves taking out all snapshots belonging to the red peak within the corresponding window. This forms a 'purified' set of the sample data. Since

the red and blue peaks in all of the plots in Fig. 1 are sharp and well separated, the width of the window for either the blue or the red peak can be defined as the distance between these two peaks along the horizontal axis. That is, the window for each peak is defined as beginning at the position 'peak centre minus half width' and ending at the position 'peak centre plus half width'. The resulting numbers of snapshots in the red peak and the blue peak for different samples are listed in Table 3. As can be seen, there are still impurities (native snapshots) in the 'purified' data of the Gd derivative (the red peak). This may be caused by the nature of SFX diffraction. A Bragg reflection under the simulation conditions should be a small sphere in reciprocal space. Each snapshot is produced from a not necessarily identical part of the reflection which intersects with the Ewald sphere. This leads to positional errors in reciprocal space and hence to errors in unit-cell parameters. However, the percentages of impurities after purification are about 100 times smaller than that of the original data listed in Table 1. In more complicated cases there may be more than one species of impurity and the peak of the main component

Table 4
Bijvoet ratio and correctness of the sign of ΔF for different samples before and after post-experimental purification.

Sample No.	Impurities (%)	Bijvoet ratio† (%)		Correctness of sign of ΔF‡ (%)	
		Original data	Purified data	Original data	Purified data
1	0	9.86	—	78.27	—
2	1	9.84	9.85	78.18	78.18
3	2	9.79	9.87	78.20	78.20
4	3	9.74	9.86	78.09	78.14
5	4	9.63	9.84	77.92	78.10
6	5	9.52	9.85	77.71	78.19
7	6	9.44	9.86	77.83	78.21
8	7	9.34	9.84	77.53	78.21
9	8	9.24	9.87	77.65	78.26
10	9	9.14	9.85	77.34	78.25
11	10	9.12	9.87	77.06	78.29
12	11	8.99	9.87	77.00	78.20
13	12	8.93	9.88	77.02	78.25
14	13	8.88	9.89	76.84	78.19
15	14	8.76	9.86	76.64	78.24
16	15	8.69	9.85	76.43	78.21

† Bijvoet ratio = $\langle|F^+ - F^-|\rangle/\frac{1}{2}\langle F^+ + F^-\rangle$, where $\langle\ldots\rangle$ denotes the average over reciprocal space. ‡ $\Delta F = F^+ - F^-$.

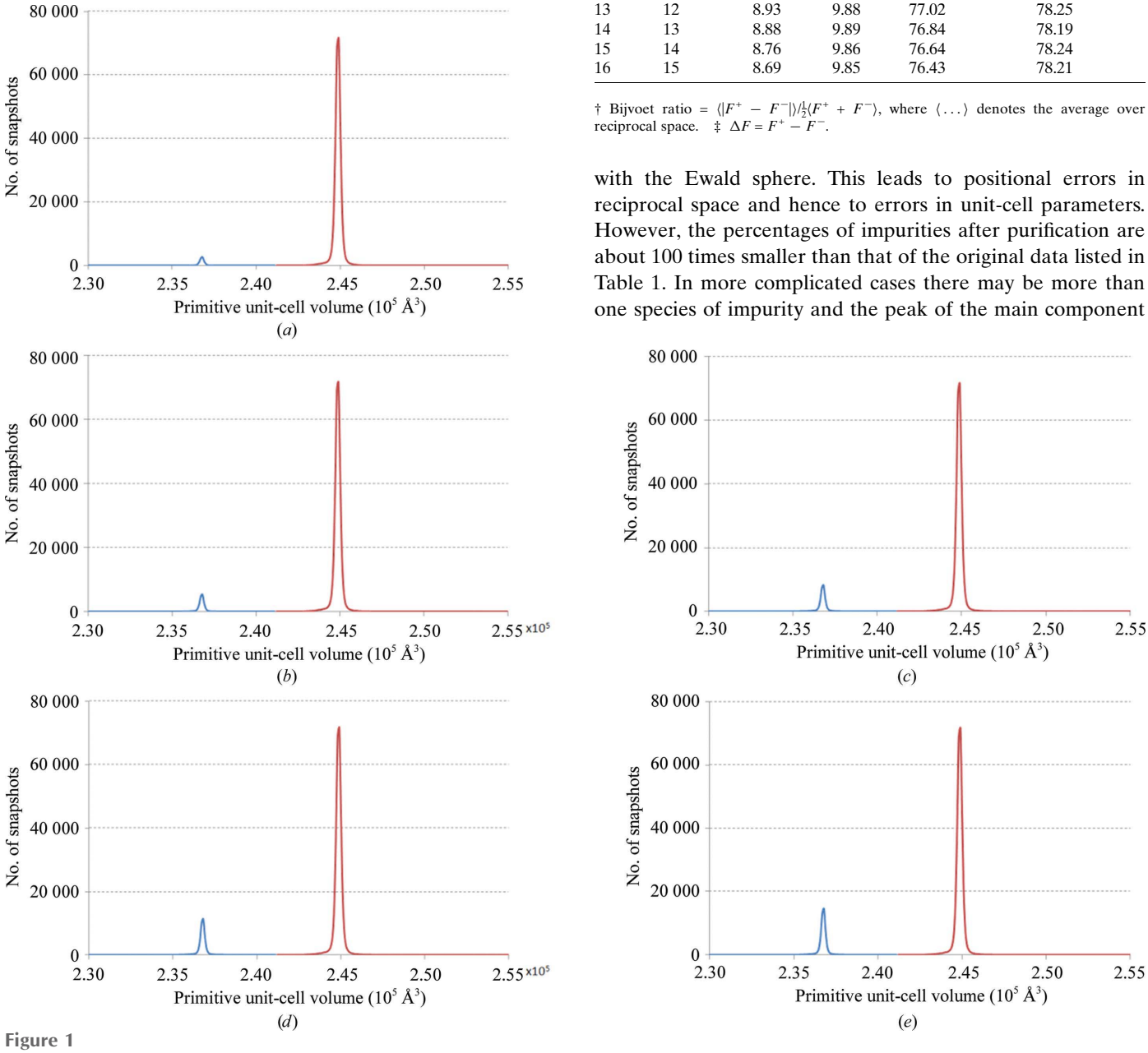

Figure 1
Number of diffraction snapshots as a function of primitive unit-cell volume plotted for different samples. (*a*), (*b*), (*c*), (*d*) and (*e*) are the results from sample Nos. 3, 6, 9, 12 and 15, respectively (see Table 1).

Table 5
The best CC_{all} and CC_{weak} within 10 000 trials of *SHELXD* and the accuracy of the resultant heavy atoms for different diffraction data sets.

	CC_{all} (%)		CC_{weak} (%)		Averaged deviation† (Å) of Gd atoms in the heavy-atom substructure	
Impurities (%)	Original data	Purified data	Original data	Purified data	Original data	Purified data
0	41.38	—	24.69	—	0.061	—
1	41.53	41.29	24.82	23.82	0.065	0.061
2	41.71	41.22	24.66	24.05	0.063	0.055
3	41.20	41.10	23.73	23.72	0.055	0.064
4	41.13	41.24	23.55	23.71	0.061	0.070
5	41.21	41.21	23.75	23.89	0.063	0.065
6	41.47	41.30	23.83	23.95	0.051	0.060
7	41.21	41.17	23.88	23.61	0.050	0.052
8	41.09	41.33	23.22	23.97	0.067	0.064
9	40.68	41.37	23.09	23.21	0.061	0.066
10	40.48	41.39	23.16	23.62	0.075	0.061
11	40.10	41.33	22.76	24.09	0.074	0.066
12	39.86	41.32	22.50	23.94	0.069	0.067
13	39.66	41.38	22.19	23.66	0.064	0.066
14	39.28	41.40	22.01	24.06	0.077	0.069
15	39.01	41.21	20.75	23.81	0.077	0.068

† Calculated against the two Gd atoms in the final structure of the Gd derivative of lysozyme (PDB entry 4n5r).

Table 6
Summarized results of structure solution for different diffraction data sets.

	Phase error (°)				No. of residues with ΔC^{α}† < 1 Å (average deviation of C^{α} in Å)	
	Phaser		*DM*		*ARP/wARP*	
Impurities (%)	Original data	Purified data	Original data	Purified data	Original data	Purified data
0	51.9	—	48.7	—	126 (0.086)	—
1	51.9	52.0	48.8	47.9	126 (0.080)	126 (0.077)
2	51.8	51.9	48.9	48.3	126 (0.082)	126 (0.079)
3	52.1	52.0	48.8	48.1	105 (0.112)	126 (0.079)
4	52.2	51.8	49.1	48.1	126 (0.082)	126 (0.087)
5	52.5	52.0	48.8	48.3	126 (0.084)	126 (0.090)
6	52.6	52.0	49.2	48.4	93 (0.170)	126 (0.088)
7	52.9	51.9	49.3	48.6	76 (0.224)	127 (0.090)
8	53.0	51.8	49.6	48.3	84 (0.182)	126 (0.087)
9	53.3	52.0	49.6	48.3	45 (0.346)	126 (0.081)
10	53.4	52.1	49.9	48.5	75 (0.217)	127 (0.081)
11	53.5	51.8	50.1	48.4	56 (0.264)	127 (0.089)
12	53.7	51.9	50.0	48.6	57 (0.304)	126 (0.087)
13	54.0	52.0	50.9	48.4	55 (0.301)	127 (0.084)
14	54.2	52.0	51.1	48.1	80 (0.235)	126 (0.087)
15	54.3	51.9	51.4	48.2	49 (0.375)	126 (0.081)

† ΔC^{α} is the positional deviation of C^{α} atoms in the built model from those of the original structure.

may overlap seriously with those of impurities. In principle, this may be handled by just narrowing the width of the window that contains the highest peak. The gain from post-experimental purification can first be seen from the changes in the Bijvoet ratio and the correctness of the sign of the Bijvoet difference ΔF (= $F^{+} - F^{-}$) before and after purification, which are listed in Table 4. For the original data, the Bijvoet ratio and the correctness of the sign of the Bijvoet difference decrease as the percentage of impurities increases. On the other hand, for the purified data the Bijvoet ratio and the correctness of the sign of the Bijvoet difference remain almost constant and are nearly the same as those of the pure sample (No. 1). While both the Bijvoet ratio and the correctness of the sign of the Bijvoet difference in Table 4 can be one of the quality indicators for a set of diffraction data, they are all overall averaged values. Hence, they do not directly affect the results of structure analysis. What directly influences the structure-analysis results are the changes in the Bijvoet ratio or Bijvoet difference of individual reflections. In the following, the effect of impurities in solving the substructure of anomalous scatterers (Gd atoms) will be given in §4, while the effect

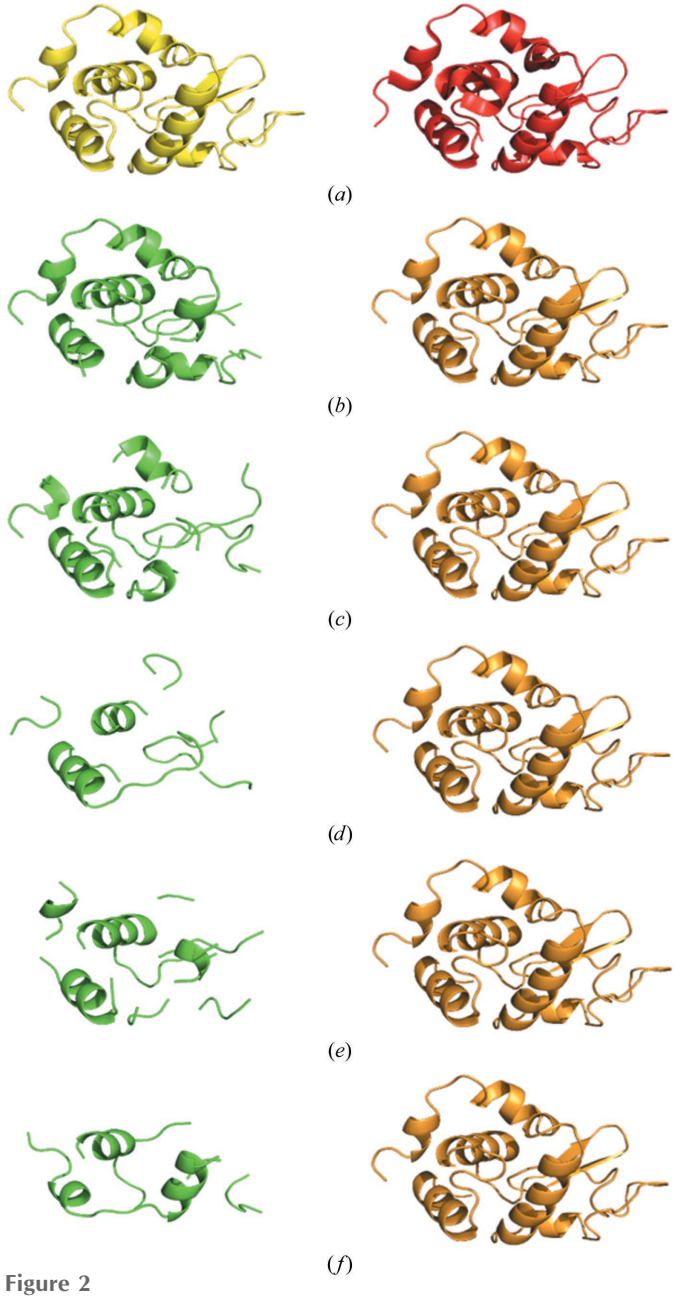

Figure 2
Cartoon structure models plotted by *PyMOL* (DeLano, 2002). (*a*) Left, the resultant model from sample No. 1 (the pure Gd-derivative sample); right, the final structure model (PDB entry 4n5r). (*b*), (*c*), (*d*), (*e*) and (*f*) result from sample Nos. 3, 6, 9, 12 and 15, respectively. Left, using original data; right, using purified data.

in solving the whole protein structure will be described in §5. We now have, for each of sample Nos. 2–16, a 'purified' subset within the original set. All original and purified data sets together with the 'pure' Gd-derivative data (sample No. 1) were treated separately to extract diffraction intensities by the procedure of Zhang, Li *et al.* (2014). The resultant diffraction data sets were then used separately in all the following tests.

4. Test on solving the heavy-atom substructure

Each diffraction data set was treated separately by *SHELXC/D* (Sheldrick, 2010) running with default settings. The only exception was that the number of random-phase trials of *SHELXD* was set to 10 000. Usually, 1000 or even 100 trials of *SHELXD* are sufficient to solve the heavy-atom substructure of proteins. The reason for the use of 10 000 trials here is to eliminate strong fluctuations that are not related to SAD signals. The best CC_{all}, CC_{weak} and the accuracy of the associated resultant heavy atoms from each 10 000 trials of *SHELXD* are listed in Table 5. It turns out that for the original data sets, apart from some small fluctuations, both CC_{all} and CC_{weak} decrease as the percentage of impurities increases, while for 'purified' data sets the values of CC_{all} and CC_{weak} are close to those for the pure sample. The same holds for the accuracy of Gd atoms, except that there are stronger fluctuations that are not related to SAD signals. On the whole, the influence of impurities on the solution of the Gd substructure is obvious but not serious.

5. Test on SAD-phasing structure solution of the Gd derivative of lysozyme

Each set of diffraction intensities was passed through the automatic *de novo* SAD-phasing structure-solution process based on the heavy-atom substructure obtained in the previous section. The process was implemented by a combination of *Phaser* (McCoy *et al.*, 2007), *DM* (Cowtan, 1994), *ARP/wARP* (Perrakis *et al.*, 1999) and *REFMAC* (Murshudov

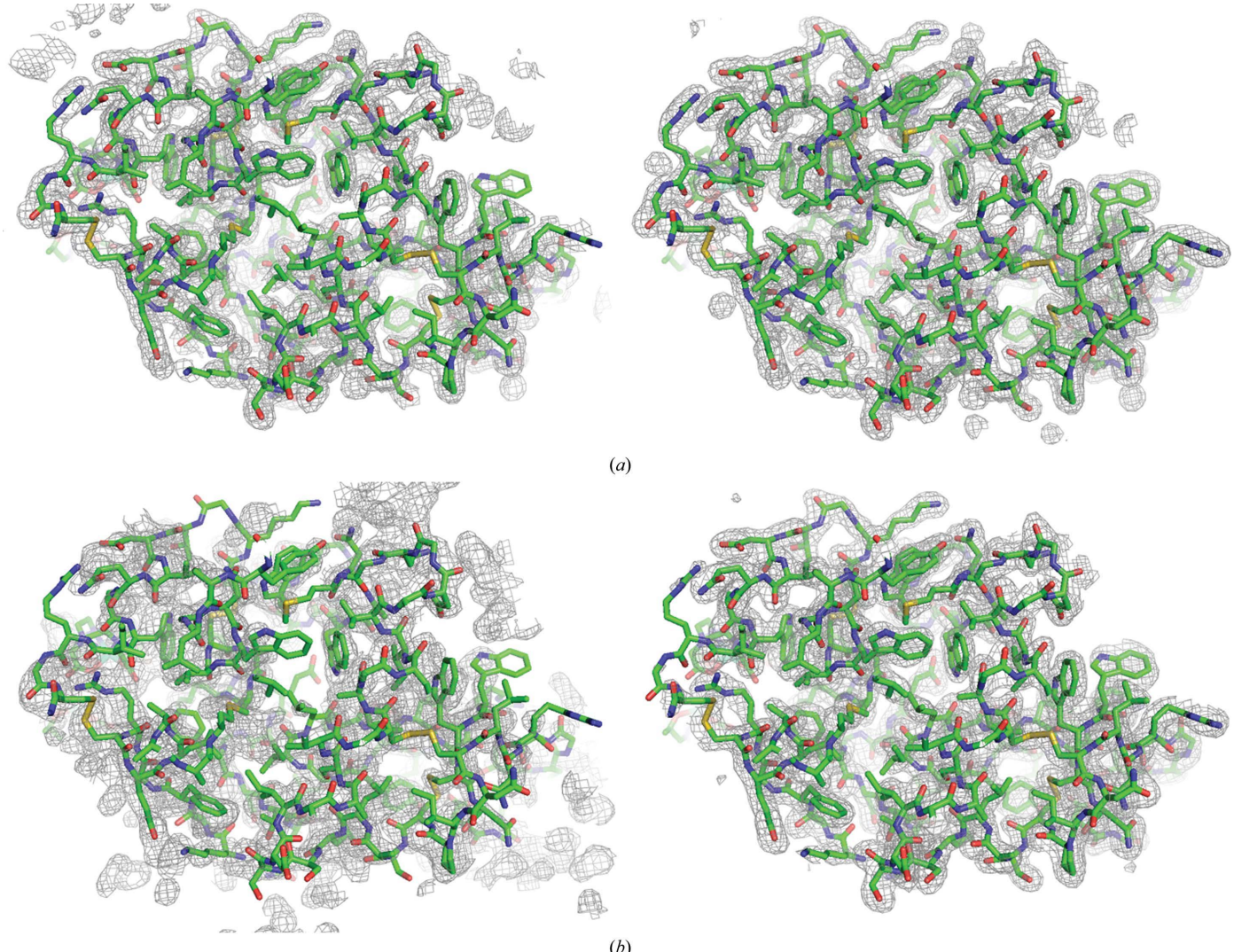

Figure 3
The same portion of electron-density maps calculated from different sample data plotted by *PyMOL* (DeLano, 2002). (*a*) From the data containing 2% impurities; (*b*) from the data containing 15% impurities; Left, before purification; right, after purification. Electron-density contours are plotted at 1σ.

et al., 2011). The results are summarized in Table 6. It was found that for the original data sets the phase error resulting from either *Phaser* or *DM* increases as the percentage of impurities increases, apart from some small fluctuations, while the number of residues built in the model and the accuracy of the C^{α} atoms decreases rapidly. On the other hand, the results from all 'purified' data sets are similar to that for the pure sample, without an evident dependence on the percentage of impurities. This can be observed intuitively in Fig. 2, where cartoon structure models are shown for diffraction data with different impurity concentrations. In Fig. 3 the same portion of the electron-density map from different samples is shown before and after purification. As can be seen, for the sample containing 15% impurities (Fig. 3*b*) the quality of the electron-density map is dramatically improved after purification. Even for the sample containing only 2% impurities (Fig. 3*a*) the improvement after purification is still obvious. By comparing the last two columns in Table 6 with those in Table 5, it is clear that the change in impurity concentration has a much stronger effect on the averaged deviation of C^{α} atoms (in solving the protein structure) than on that of Gd atoms (in solving the heavy-atom substructure). For C^{α} atoms the averaged deviation changes from 0.09 to 0.38 Å when the percentage of impurities changes from 0 to 15%, while under the same conditions the averaged deviation of Gd atoms changes only from 0.06 to 0.08 Å, which is about 15 times smaller than that for C^{α} atoms. This can be explained as follows: in the solution of the heavy-atom substructure changes in the impurity concentration directly affect the magnitudes $|F^{+} - F^{-}|$, while in the SAD-phasing structure solution of the protein changes in the impurity concentration directly affect the phases. As has been well known for decades, the quality of electron-density maps is much more sensitive to phases than to magnitudes.

6. Concluding remarks

The present test was performed assuming that an X-ray free-electron laser source was used. However, it is obvious that similar results can also be obtained with third-generation synchrotron sources. Gd-derivative lysozyme SFX data with native data as impurities mimic the common scenario of using a huge number of micrometre-sized crystals as the diffraction sample. Our test demonstrated that a small amount of sample impurities could reduce the quality of the diffraction intensities, making SAD phasing with SFX data more difficult. On the other hand, post-experimental purification is capable of efficiently removing sample impurities and substantially improving the structure-solution results.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 11204364). HF would like to thank the referees, whose suggestions improved the manuscript.

References

Barends, T. R. M., Foucar, L., Botha, S., Doak, R. B., Shoeman, R. L., Nass, K., Koglin, J. E., Williams, G. J., Boutet, S., Messerschmidt, M. & Schlichting, I. (2014). *Nature (London)*, **505**, 244–247.
Botha, S., Nass, K., Barends, T. R. M., Kabsch, W., Latz, B., Dworkowski, F., Foucar, L., Panepucci, E., Wang, M., Shoeman, R. L., Schlichting, I. & Doak, R. B. (2015). *Acta Cryst.* D**71**, 387–397.
Boutet, S. *et al.* (2012). *Science*, **337**, 362–364.
Chapman, H. N. *et al.* (2011). *Nature (London)*, **470**, 73–77.
Cowtan, K. (1994). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **31**, 34–38.
DeLano, W. L. (2002). *PyMOL*. http://www.pymol.org.
Gati, C., Bourenkov, G., Klinge, M., Rehders, D., Stellato, F., Oberthür, D., Yefanov, O., Sommer, B. P., Mogk, S., Duszenko, M., Betzel, C., Schneider, T. R., Chapman, H. N. & Redecke, L. (2014). *IUCrJ*, **1**, 87–94.
Lu, F., Li, S., Jiang, Y., Jiang, J., Fan, H., Lu, G., Deng, D., Dang, S., Zhang, X., Wang, J. & Yan, N. (2011). *Nature (London)*, **472**, 243–246.
McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.
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.
Nogly, P. *et al.* (2015). *IUCrJ*, **2**, 168–176.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Sheldrick, G. M. (2010). *Acta Cryst.* D**66**, 479–485.
Stellato, F. *et al.* (2014). *IUCrJ*, **1**, 204–212.
White, T. A., Kirian, R. A., Martin, A. V., Aquila, A., Nass, K., Barty, A. & Chapman, H. N. (2012). *J. Appl. Cryst.* **45**, 335–341.
Yamashita, K. *et al.* (2015). *Sci. Rep.* **5**, 14017.
Yao, D., Cherney, M. & Cygler, M. (2014). *Acta Cryst.* D**70**, 436–441.
Yao, D., Zhang, T., He, Y., Han, P., Cherney, M., Gu, Y., Cygler, M. & Fan, H. (2014). *Acta Cryst.* D**70**, 2686–2691.
Zhang, T., Li, Y. & Wu, L. (2014). *Acta Cryst.* A**70**, 670–676.
Zhang, T., Yao, D., Wang, J., Gu, Y. & Fan, H. (2015). *Acta Cryst.* D**71**, 2513–2518.