BIOLOGICAL CRYSTALLOGRAPHY
ISSN 1399-0047

Serial crystallographic analysis of protein isomorphous replacement data from a mixture of native and derivative microcrystals

Tao Zhang,[a] Deqiang Yao,[b] Jiawei Wang,[c] Yuanxin Gu[a] and Haifu Fan[a]*

[a]Institute of Physics, Chinese Academy of Sciences, Beijing 100190, People's Republic of China, [b]Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, People's Republic of China, and [c]School of Life Sciences, Tsinghua University, Beijing 100084, People's Republic of China. *Correspondence e-mail: fan@iphy.ac.cn

Received 22 April 2015
Accepted 27 August 2015

Edited by Q. Hao, University of Hong Kong

Keywords: serial crystallography; multiphase diffraction; isomorphous replacement.

A post-experimental identification/purification procedure similar to that described in Zhang *et al.* [(2015), *IUCrJ*, **2**, 322–326] has been proposed for use in the treatment of multiphase protein serial crystallography (SX) diffraction snapshots. As a proof of concept, the procedure was tested using theoretical serial femtosecond crystallography (SFX) data from a mixture containing native and derivatized crystals of a protein. Two known proteins were taken as examples. Multiphase diffraction snapshots were subjected to two rounds of indexing using the program *CrystFEL* [White *et al.* (2012). *J. Appl. Cryst.* **45**, 335–341]. In the first round, an *ab initio* indexing was performed to derive a set of approximate primitive unit-cell parameters, which are roughly the average of those from the native protein and the derivative. These parameters were then used in a second round of indexing as input to *CrystFEL*. The results were then used to separate the diffraction snapshots into two subsets corresponding to the native and the derivative. For each test sample, integration of the two subsets of snapshots separately led to two sets of three-dimensional diffraction intensities, one belonging to the native and the other to the derivative. Based on these two sets of intensities, a conventional single isomorphous replacement (SIR) procedure solved the structure easily.

1. Introduction

The 'diffraction before destruction' approach of serial femtosecond crystallography (SFX; Chapman *et al.*, 2011) was proposed for use with a hard X-ray free-electron laser (XFEL) to avoid radiation damage to the sample crystals during diffraction data collection. The technique has recently been extended to use with third-generation synchrotron radiation (Gati *et al.*, 2014; Stellato *et al.*, 2014; Botha *et al.*, 2015; Nogly *et al.*, 2015). Serial crystallography (SX) opens up new possibilities for solving the crystal structures of biological macromolecules. (i) Complicated protein structures can be solved using micrometre-sized polycrystalline samples. Such crystal dimensions are about ten times smaller than those used in conventional single-crystal protein structure determination with third-generation synchrotron sources. (ii) Diffraction data collection can be performed at room temperature rather than in a cryocooling environment. Cryoprotection can induce a significant increase in crystal mosaicity (Guha *et al.*, 2012), while room temperature can provide an environment close to the living conditions of proteins. (iii) It is possible to study *in vivo*-grown protein microcrystals (Koopmann *et al.*, 2012) even without taking them out of the growing cells (Axford *et al.*, 2014). (iv) Time-resolved structure studies are enabled (Aquila *et al.*, 2012; Neutze & Moffat, 2012; Spence *et al.*, 2012; Kupitz *et al.*, 2014).

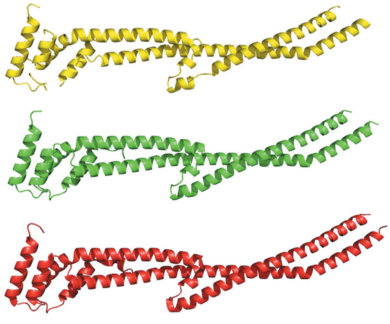

Table 1
Summarized crystallographic data for the proteins LegC3N and NAT/NCS2.

	LegC3N		NAT/NCS2	
	Native	Hg derivative	Native	Hg derivative
Space group	$P2_12_12$		$P6_422$	
Unit-cell parameters				
a (Å)	108.874	107.039	96.760	97.227
b (Å)	150.246	149.507	96.760	97.227
c (Å)	24.240	24.207	251.950	252.430
γ (°)			120	120
Unit-cell volume (Å^3)	3.965×10^5	3.874×10^5	2.043×10^6	2.067×10^6
Heavy atoms per asymmetric unit	3 × Hg		4 × Hg	
No. of residues per asymmetric unit	367		429	
Reference (PDB code)	Yao *et al.*, 2014 (4um6)		Lu *et al.*, 2011 (3qe7)	

Table 2
Conditions for simulating calculations of diffraction snapshots by SFX.

Photoenergy (eV)	1.24×10^4
Pulse fluence (photons μm^{-2})	7.0735×10^{11}
Beam bandwidth	0.001
Beam divergence (rad)	0.001
Pixels on detector	1456 × 1456
Pixel size (μm)	110
Sample-to-detector distance (cm)	20
Dimension (nm) of crystal grains	500–5000
Atomic parameters of the sample proteins	Taken from PDB entries 4um6 for LegC3N and 3qe7 for NAT/NCS2
Averaged Poisson noise (applied by default in *CrystFEL*)	6.6% for LegC3N, 7.1% for NAT/NCS2

Table 3
Summarized results of SFX diffraction-snapshot simulation.

	LegC3N		NAT/NCS2	
	Native	Hg derivative	Native	Hg derivative
No. of simulated snapshots	150000	100000	100000	100000
No. of indexed snapshots†	135028	92448	84678	90305
Resolution (Å)	61.287–1.985		83.983–2.604	
Total No. of reflections	25815		21258	
Multiplicity (overall)	576	404	1001	1064
R_{split} (overall)	0.266	0.303	0.182	0.152

† Counted for the second round of indexing.

SX is an emerging technique, in which there are problems to be solved and new applications to be explored. Recently, efforts have been made towards resolving the indexing ambiguity and improving the process of intensity integration (Brehm & Diederichs, 2014; Kabsch, 2014; Hattne *et al.*, 2014; Zeldin *et al.*, 2015; Sauter, 2015). These could help to strengthen the power of the technique. Apart from data processing, there are many phasing methods which are very successful in conventional single-crystal protein structure analysis. In principle, these methods should also be applicable in SX, and some of them have been tested in SX. Barends *et al.* (2014) reported a test of gadolinium SAD phasing with the known protein lysozyme. Botha *et al.* (2015) demonstrated multiple isomorphous replacement with anomalous scattering (MIRAS) and single isomorphous replacement with anomalous scattering (SIRAS) phasing using native, iodide and gold derivatives of lysozyme. However, to date only the molecular-replacement method has been successful in SX for solving the structures of originally unknown proteins. The main obstacle in practice may be the quality of diffraction intensities. Apart from the problem of the partial diffraction recorded for each Bragg reflection and the inaccuracy of integration methods, the heterogeneity of the microcrystals used in SX experiments also degrades the quality of diffraction intensities (Sauter, 2015). We have found that a small amount of impurities in the heavy-atom derivative sample can strongly affect the results of SAD phasing in SX (Zhang, Jin *et al.*, 2014). In this paper, an SX SIR/MIR phasing procedure will be described. It is a combination of the conventional protein SIR/MIR phasing technique with a post-experimental identification/purification process dedicated to the treatment of SX diffraction snapshots from a multiphase mixture sample.

2. Test samples and simulation

For the present test, we prepared two test samples. Both are diffraction snapshots of an SIR mixture (containing snapshots from both the native and a derivative). One of the two samples is associated with the protein LegC3N (Yao *et al.*, 2014), while the other is related to NAT/NCS2 (Lu *et al.*, 2011). Crystallographic parameters for the two proteins are summarized in Table 1. The structures of both proteins were originally solved by the SIRAS method. The simulation was based on the experimental unit-cell parameters and the final structure models from the PDB entries (see Table 1). The *CrystFEL* program suite (White *et al.*, 2012) was used for simulation under the conditions listed in Table 2. For each set of mixed SIR data, diffraction snapshots of the native and derivative were first calculated separately and then mixed randomly in a ratio of 1.5:1 for LegC3N and in a ratio of 1:1 for NAT/NCS2. The results of the simulation are summarized in Table 3.

3. Identification and diffraction-intensity extraction

At the beginning of the present test, we assumed it to already be known that the two sets of test diffraction snapshots are from the proteins LegC3N and NAT/NCS2; both contain mainly diffraction snapshots from the native and the derivative without knowing the unit-cell parameters. Indexing was performed using the program *indexamajig* in *CrystFEL* running with default settings. The number of indexable snapshots was plotted against the primitive unit-cell volume as shown in Figs. 1 and 2 for the proteins LegC3N and NAT/NCS2, respectively. Figs. 1(*a*) and 2(*a*) show the full range distribution; Figs. 1(*b*) and 2(*b*) show an enlarged portion covering only the highest peak in Figs. 1(*a*) and 2(*a*). One dominating peak and a number of much smaller peaks can be

found in Figs. 1(*a*) and 2(*a*). The most prominent small peaks are on the right-hand side of the dominating peak and have a primitive unit-cell volume N times (where N is an integer) larger than that of the dominating peak. Most probably, they belong to the same crystallographic phase as the dominating peak. Other small peaks may be owing to errors. In Figs. 1(*b*) and 2(*b*) it is shown that the dominating peaks in Figs. 1(*a*) and 2(*a*) are now split into two peaks with unit-cell parameters close to each other. This indicates that both examples mainly consist of two components, *i.e.* the native protein and the derivative. On the other hand, the peaks in Figs. 1(*b*) and 2(*b*) cover only a small portion of the whole snapshots from LegC3N and NAT/NCS2, respectively. The separation directly from the two figures will be far from complete and subsequent structure solution based on it may be problematic. Hence, a procedure based on two rounds of indexing is used in the present work. The only task of the first round is to derive an approximate set of primitive unit-cell parameters. For this purpose, it is not necessary to use the whole set of snapshots. 5000 snapshots were randomly selected for *ab initio* indexing by the program *indexamajig* running under the default control of *CrystFEL*. This led to a native–derivative averaged primitive unit cell of the mixed SIR data for LegC3N with $a = 108.21$, $b = 149.87$, $c = 24.21$ Å, $\alpha = 89.97$, $\beta = 89.94$, $\gamma = 89.97^\circ$ and for NAT/NCS2 with $a = 97.13$, $b = 97.41$, $c = 252.11$ Å, $\alpha = 89.92$, $\beta = 89.97$, $\gamma = 60.09^\circ$.

In the second round, all individual snapshots of each protein were subjected to indexing. The primitive unit-cell parameters obtained in the first round were input to the program *CrystFEL*, which was again running under default control. Distributions of the number of indexable snapshots against the primitive unit-cell volume from the second round of indexing are plotted in Figs. 3 and 4 for LegC3N and for NAT/NCS2, respectively. Here, only the enlarged portions corresponding to those of Figs. 1(*b*) or 2(*b*), respectively, are given. In comparison with Figs. 1(*b*) or 2(*b*), many more snapshots are now included in Figs. 3 and 4 and the separation of the two components will accordingly be much more complete. In the latter two figures, the peak on the left (with a smaller primitive volume) is denoted in blue while that on the right is in red. Obviously, of the blue and red peaks, one should correspond to the native and the other to the derivative. Accordingly, for each protein we can extract two subsets of snapshots corresponding to the native and the derivative. In order to extract a set of snapshots corresponding to a peak in Figs. 3 or 4, we need to specify the position of its maximum and define the length of the baseline (coincident with the abscissa of Figs. 3 or 4 with its midpoint set at the maximum position of

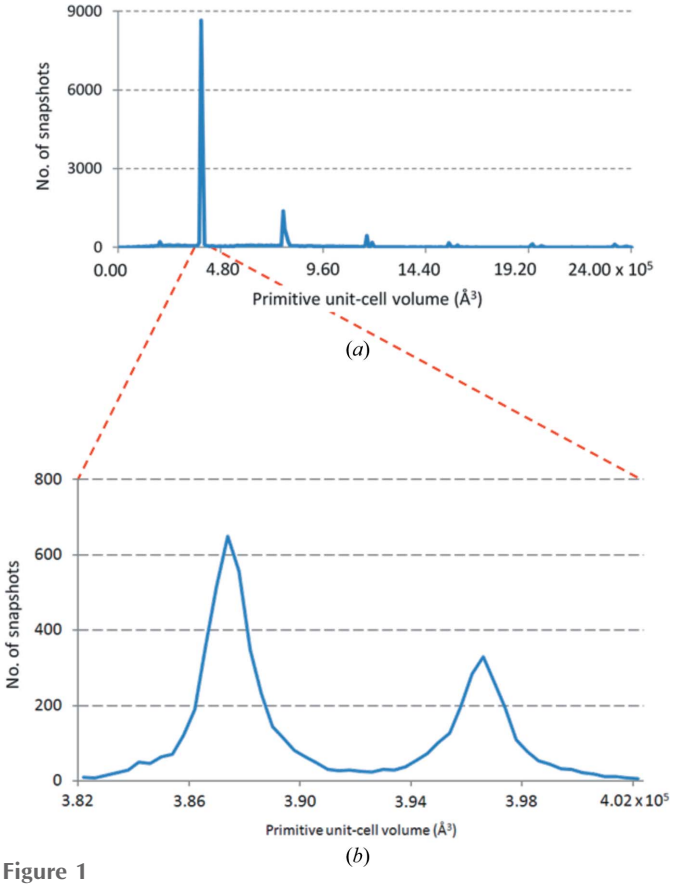

Figure 1
Result of the preliminary indexing for LegC3N: distribution of the number of indexable diffraction snapshots against the primitive unit-cell volume. (*a*) Full range distribution; (*b*) enlarged portion covering only the highest peak in (*a*).

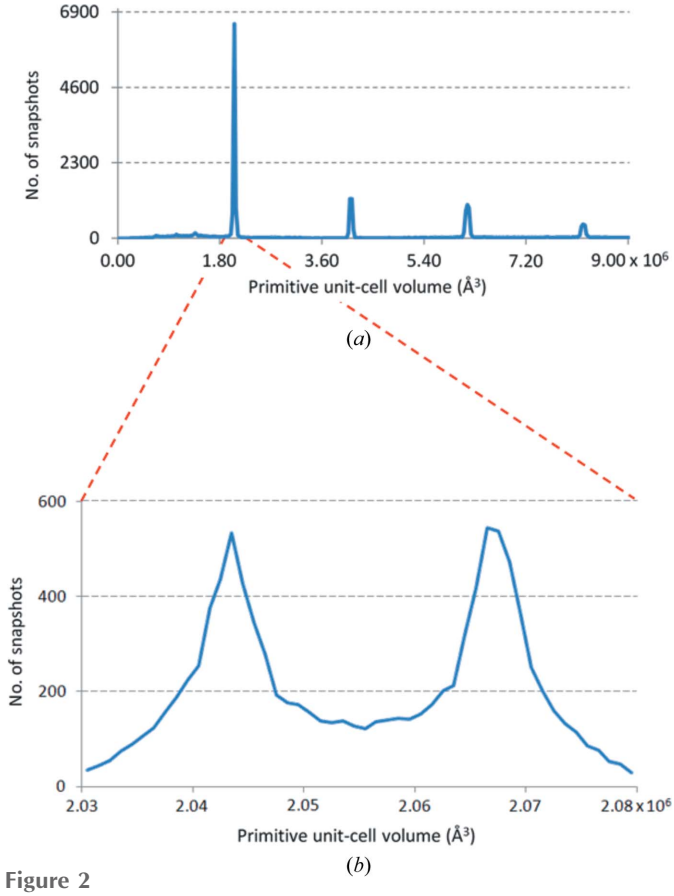

Figure 2
Result of the preliminary indexing for NAT/NCS2: distribution of the number of indexable diffraction snapshots against the primitive unit-cell volume. (*a*) Full range distribution; (*b*) enlarged portion covering only the highest peak in (*a*).

Table 4
CC and CC_{weak} values resulting from *SHELXD* for different native/derivative assignments.

	LegC3N		NAT/NCS2	
Native/derivative assignment	CC	CC_{weak}	CC	CC_{weak}
Native, red; derivative, blue	36.79	24.48	24.36	11.69
Native, blue; derivative, red	29.23	19.82	49.66	36.40

the peak). All snapshots with the primitive unit-cell volume falling on the baseline should be extracted to form a subset with the averaged primitive unit-cell volume equal to that of the maximum position. A post-experimental purification of a particular component can now simply be performed by shortening the length of the baseline. In the present test, since no impurities have been assumed in the test sample, no improvements in data quality could be found by shortening the baseline of either the native peak or the derivative peak. Hence, we just set the length equal to the distance between the maxima of the red and the blue peaks. By extracting snapshots from the red and blue peaks in Figs. 3 and 4, we obtained four sets of diffraction snapshots. Intensity extraction was performed for each of the four snapshot sets separately using the procedure of Zhang, Li *et al.* (2014) instead of Monte Carlo integration (Kirian *et al.*, 2010). The former method

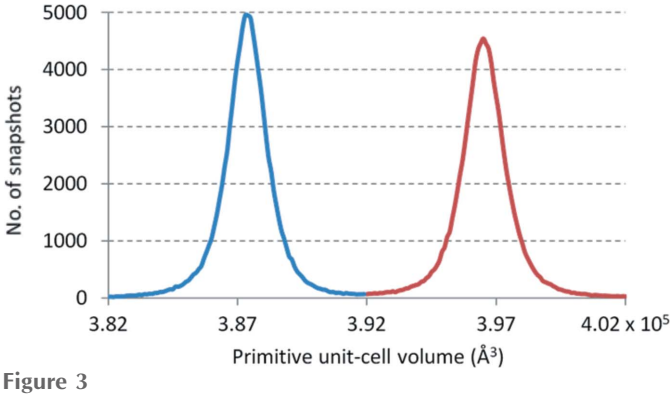

Figure 3
Result of the second round of indexing for LegC3N: distribution of the number of indexable diffraction snapshots against the primitive unit-cell volume. Only the enlarged portion corresponding to Fig. 1(*b*) is shown.

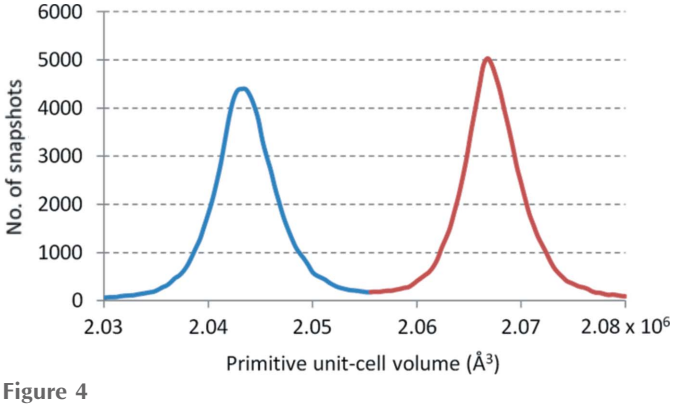

Figure 4
Result of the second round of indexing for NAT/NCS2: distribution of the number of indexable diffraction snapshots against the primitive unit-cell volume. Only the enlarged portion corresponding to Fig. 2(*b*) is shown.

Table 5
Resultant structure models from *SHELXC*/*D*/*E* and *ARP*/*wARP*.

		No. of residues				
Protein	Program	Built	Placed	$\Delta C^{\alpha} < 1$ Å†	R	R_{free}
LegC3N	*SHELXC*/*D*/*E*	234‡	—	213‡	—	—
	ARP/*wARP*	260	260	258	0.23	0.28
NAT/NCS2	*SHELXC*/*D*/*E*	316‡	—	199‡	—	—
	ARP/*wARP*	401	391	379	0.25	0.28

† ΔC^{α} is the positional deviation of C^{α} atoms in the built model from that of the final structure. ‡ From the polyalanine model.

gave better results and was able to use fewer snapshots in the present test. After diffraction-intensity extraction, the four sets of intensity data were treated separately with *XPREP* (http://shelx.uni-ac.gwdg.de/tutorial/english/). This led to four sets (two belonging to LegC3N and the other two belonging to NAT/NCS2) of diffraction data with space groups and crystallographic unit cells matching those listed in Table 1. Up to this point, we have diffraction data for two SIR pairs for the proteins LegC3N and NAT/NCS2. However, in each SIR pair we do not know whether the red peak corresponds to the native or to the derivative. The same holds for the blue peak. This problem remains to be solved in the following section.

4. Structure determination

Diffraction data for the two SIR pairs obtained in the previous section were used separately to solve the structures of LegC3N and NAT/NCS2. *SHELXC*/*D*/*E* (Sheldrick, 2010) as implemented *via* the *HKL2MAP* GUI (Pape & Schneider, 2004) were used for diffraction-intensity normalization, native–derivative discrimination, heavy-atom substructure determination, SIR phasing, density modification and

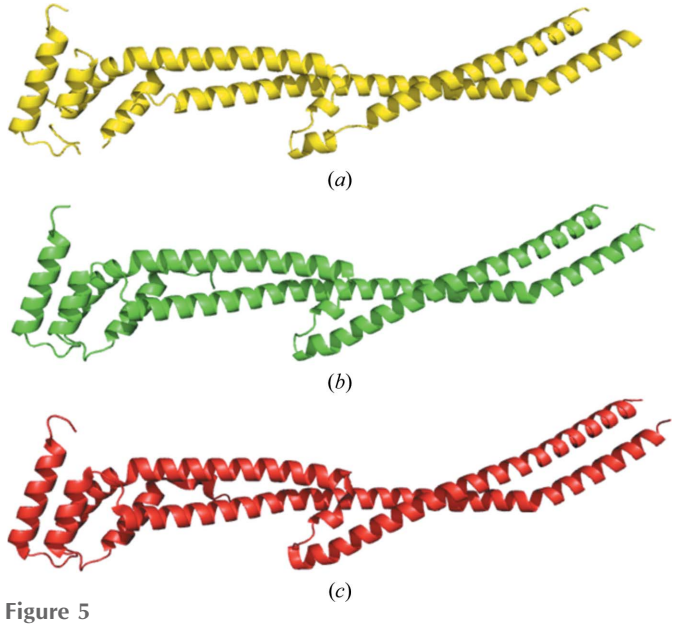

Figure 5
Cartoon structure models of LegC3N. (*a*) *SHELXC*/*D*/*E* polyalanine model derived from the simulated SFX SIR data. (*b*) *ARP*/*wARP* model based on (*a*). (*c*) The final model from PDB entry 4um6.

(*a*) (*b*) (*c*)

Figure 6
Cartoon structure models of NAT/NCS2. (*a*) *SHELXC*/*D*/*E* polyalanine model derived from the simulated SFX SIR data. (*b*) *ARP*/*wARP* model based on (*a*). (*c*) The final model from PDB entry 3qe7.

polyalanine model building. For resolving the native–derivative ambiguity and locating the heavy-atom sites, *SHELXC*/*D* were run twice. In the first run, diffraction data from the red peaks were assigned as from the native, while those from the blue peaks were assigned as from the derivative. In the second run the assignment was inverted. The different assignments should affect the normalization in *SHELXC*, leading to different results from *SHELXD*. After 1000 trials for each run of *SHELXD*, the best (largest) CC and CC_{weak} values are listed in Table 4. It is expected that the correct assignment should lead to larger CC and CC_{weak} values and the correct heavy-atom substructure should be that associated with the largest CC and CC_{weak}. Hence, in view of Table 4, we concluded that the red peak in Fig. 3(*b*) is from native LegC3N, while the blue peak in Fig. 4(*b*) is from native NAT/NCS2. Accordingly, we also obtained the heavy-atom substructure from the *SHELXD* output for the derivatives of LegC3N and of NAT/NCS2. The two heavy-atom substructures were then passed onto *SHELXE* separately for SIR phasing, density modification and polyalanine model building. All steps were run using the program defaults. Finally, *ARP*/*wARP* (Langer *et al.*, 2008), also running with default settings, was used to extend and refine the resultant polyalanine models from *SHELXE*. This led to almost complete structures of LegC3N and NAT/NCS2. Structure-analysis results are summarized in Table 5. Cartoon structure models plotted by *PyMOL* (DeLano, 2002) are shown in Figs. 5 and 6 for LegC3N and NAT/NCS2, respectively.

5. Concluding remarks

Unlike single-crystal diffraction experiments, an SX experiment requires tens of thousands of crystal grains. It would be difficult to ensure the purity of such a sample by conventional experimental treatments. However, the post-experimental identification/purification procedure may solve the problem effectively. The present test was performed assuming that an X-ray free-electron laser source is used; it is obvious that similar results can also be obtained with a synchrotron-radiation source. Although the present test was only performed with SIR data, there is no doubt that the proposed method is also applicable to MIR data. Besides, significant alternate conformations of side chains and the backbone and different configurations/conformations of multi-component macromolecular complexes might result in a variation of the unit-cell parameters. Such problems may also be tackled using the proposed method.

Acknowledgements

HF would like to thank Professor M. Cygler and Professor N. Yan for making the data for LegC3N and NAT/NCS2 available, respectively. This work was supported in part by the National Natural Science Foundation of China (Grant No. 11204364).

References

Aquila, A. *et al.* (2012). *Opt. Express*, **20**, 2706–2716.
Axford, D., Ji, X., Stuart, D. I. & Sutton, G. (2014). *Acta Cryst.* D**70**, 1435–1441.
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.
Brehm, W. & Diederichs, K. (2014). *Acta Cryst.* D**70**, 101–109.
Chapman, H. N. *et al.* (2011). *Nature (London)*, **470**, 73–77.
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.
Guha, S., Perry, S. L., Pawate, A. S. & Kenis, P. J. A. (2012). *Sens. Actuators B Chem.* **174**, 1–9.
Hattne, J. *et al.* (2014). *Nature Methods*, **11**, 545–548.
Kabsch, W. (2014). *Acta Cryst.* D**70**, 2204–2216.

Kirian, R. A., Wang, X., Weierstall, U., Schmidt, K. E., Spence, J. C. H., Hunter, M., Fromme, P., White, T., Chapman, H. N. & Holton, J. (2010). *Opt. Express*, **18**, 5713–5723.
Koopmann, R. *et al.* (2012). *Nature Methods*, **9**, 259–262.
Kupitz, C. *et al.* (2014). *Nature (London)*, **513**, 261–265.
Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. (2008). *Nature Protoc.* **3**, 1171–1179.
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.
Neutze, R. & Moffat, K. (2012). *Curr. Opin. Struct. Biol.* **22**, 651–659.
Nogly, P. *et al.* (2015). *IUCrJ*, **2**, 168–176.
Pape, T. & Schneider, T. R. (2004). *J. Appl. Cryst.* **37**, 843–844.
Sauter, N. K. (2015). *J. Synchrotron Rad.* **22**, 239–248.
Sheldrick, G. M. (2010). *Acta Cryst.* D**66**, 479–485.
Spence, J. C. H., Weierstall, U. & Chapman, H. N. (2012). *Rep. Prog. Phys.* **75**, 102601.
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.
Yao, D., Cherney, M. & Cygler, M. (2014). *Acta Cryst.* D**70**, 436–441.
Zeldin, O. B., Brewster, A. S., Hattne, J., Uervirojnangkoorn, M., Lyubimov, A. Y., Zhou, Q., Zhao, M., Weis, W. I., Sauter, N. K. & Brunger, A. T. (2015). *Acta Cryst.* D**71**, 352–356.
Zhang, T., Jin, S., Gu, Y., He, Y., Li, M. & Fan, H. (2014). *Single-Crystal Diffraction Data from Powder Samples via Serial Femtosecond Crystallography*. http://cryst.iphy.ac.cn/Publication/PDF_files/2014/XFEL_IUCr_2014_poster.pptx.
Zhang, T., Jin, S., Gu, Y., He, Y., Li, M., Li, Y. & Fan, H. (2015). *IUCrJ*, **2**, 322–326.
Zhang, T., Li, Y. & Wu, L. (2014). *Acta Cryst.* A**70**, 670–676.