Rapid Communication

Combining SAD/SIR iteration and MR iteration in partial-model extension of proteins*

Zhang Tao(张 涛)[a)b)§], Wu Li-Jie(武丽杰)[b)§], Gu Yuan-Xin(古元新)[b)†], Zheng Chao-De(郑朝德)[b)], and Fan Hai-Fu(范海福)[b)‡]

[a)] *Research Institute of Magnetic Materials, School of Physical Sciences and Technology, Lanzhou University, Lanzhou* 730000, *China*

[b)] *Beijing National Laboratory for Condensed Matter Physics and Key Laboratory of Soft Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing* 100190, *China*

(Received 28 April 2010; revised manuscript received 21 May 2010)

There are two kinds of dual-space partial-model extensions which involve the direct-method program OASIS. The first kind, named SAD/SIR iteration, uses SAD/SIR information, while the second kind, named molecular replacement (MR) iteration, does not use that information. In general, the SAD/SIR iteration is more powerful since more experimental information is used. However, in most cases when protein structures are solved with the molecular replacement method, SAD/SIR information is not available. Thus the MR iteration is particularly useful for the completion of models from molecular replacement. The SAD/SIR iteration will be automatically used in OASIS for data sets containing SAD/SIR signals, while the MR iteration will be dedicated to data sets without SAD/SIR signals. The present paper shows that for data containing SAD/SIR signals, a combination of SAD/SIR iteration and MR iteration could lead to significantly better results than that obtained from the SAD/SIR iteration alone.

Keywords: OASIS, dual-space phasing, model completion, proteins
PACC: 6110M, 8715

1. Introduction

The dual-space iterative single-wavelength anomalous diffraction or single isomorphous replacement (SAD/SIR) phasing and fragment extension was proposed in 2004,[1] a cycle of which consists of SAD/SIR phasing by OASIS,[2] density modification by DM[3,4] or RESOLVE,[5] model building and refinement by one or more of the programs RESOLVE,[6,7] ARP/wARP,[8] REFMAC[9] and AutoBuild in PHENIX.[10] This procedure, named SAD/SIR iteration in OASIS, has dramatically enhanced the SAD/SIR phasing of protein diffraction data and greatly improved the model completion when the SAD/SIR information is available.[11−13] Later, the dual-space fragment extension has been extended to protein diffraction data without SAD/SIR signals,[14] a cycle of which consists of the pseudo-MR phasing by OASIS,[2] density modification by DM[3,4] or RESOLVE,[5] model building and refinement by one or more of the programs RESOLVE,[6,7] ARP/wARP,[8] REFMAC[9] and AutoBuild in PHENIX.[10] This procedure, named molecular replacement (MR) iteration in OASIS, enhances significantly the efficiency of MR-model completion in the case that SAD/SIR signals are absent. Recently Panjikar *et al.*[15] have compared the MR iteration with the MRSAD procedure, the latter includes a process similar to the SAD/SIR iteration. They showed that while the MR iteration could effectively improve the structure model obtained from molecular replacement, the MRSAD procedure gave much better results. This conclusion is self-evident, since MRSAD uses more information (the SAD information) than the MR iteration does. On the other hand, it is interesting that Panjikar *et al.* proved the usefulness of the MR iteration in model completion when SAD/SIR signals were absent (see Table 2 in Ref. [15]). Our further investigation showed that during the final stage of model completion, even for protein diffraction data containing SAD/SIR signals,

*Work supported by the Innovation Project of the Chinese Academy of Sciences and by the National Basic Research Program of China (Grant No. 2002CB713801).
†Corresponding author. E-mail: gu@cryst.iphy.ac.cn
‡E-mail: fanhf@cryst.iphy.ac.cn
§These authors contributed equally to this work.

the MR iteration might provide better results than that from the SAD/SIR iteration, especially when SAD/SIR signals are weak and experimental errors are large. This leads to the combination of SAD/SIR iteration and MR iteration in partial-model extension of proteins.

2. Combination of SAD/SIR iteration and MR iteration

The combination is arranged as follows. For diffraction data containing reasonable SAD/SIR signals, the SAD/SIR iteration will be run first. After sufficient cycles of iteration, say 10 cycles, if the resulting structure models are not big enough, say less than 90% of the total structure, then the iteration will be changed to MR iteration taking the best resultant structure model from previous iteration cycles as input. The philosophy behind the above scheme is simple. SAD/SIR signals contain phase information, which is particularly important in the early and the middle stages of partial-model extension. This is the reason why the SAD/SIR iteration is successful (see Refs. [11]–[13]). However SAD/SIR signals also contain large experimental errors, which may bias the convergence of model completion. On the other hand, the MR iteration does not make use of SAD/SIR signals and hence is not affected by them. Besides, the algorithm in MR iteration[14] has been specially designed to avoid model bias. Thus it is reasonable to expect that the combination of SAD/SIR iteration and MR iteration will obtain better results than that from SAD/SIR iteration alone. Typical experimental diffraction data from known proteins have been used to test the above procedure.

3. Test data

Protein diffraction data used in the test are summarized in Table 1. There are three sets of SAD data and one set of SIR data. Among the SAD data sets the anomalous scatterers range from selenium, copper to sulfur. The size of the proteins ranges from 129 to 668 amino acids per asymmetric unit (AU). The high-resolution cut off of data sets is far below the 'atomic resolution' ranging from 1.9 to 2.8 Å (1 Å=0.1 nm). All samples are more or less difficult in SAD/SIR phasing. Azurin[16] has an overall completeness of only 60% and a not sufficiently high Bijvoet ratio $\langle|\Delta f|\rangle/\langle f\rangle$. Set7/9[17] is at a rather low resolution of 2.8 Å and has a low multiplicity of 3.8. TTHA1012 is a sulfur SAD data with very low Bijvoet ratio. The Rpe[18] SIR data is from one of the two derivatives, which is not as good as the other derivative (see Ref. [13]).

Table 1. Summary of test data.

protein	residues per AU	space group	x-ray wavelength/Å	heavy atoms per AU	$\langle\|\Delta F\|\rangle/\langle F\rangle$ /%	data multiplicity	resolution limit/Å	reference (PDB code)
Azurin	129	$P4_122$	0.97	1×Cu	1.45	10.0	1.9	[16] (1DYZ)
Set7/9	586	$P2_12_12_1$	0.9794	12×Se	7.03	3.8	2.8	[17] (1H3I)
TTHA1012	213	$P2_12_12_1$	2.291 (CrKα)	2×S	0.83	13.5	2.2	(2YZY)[a]
Rpe	668	R3	1.542 (CuKα)	7×Hg			2.8	[18] (1LIA)

[a] Ebihara A, Watanabe N, Yokoyama S & Kuramitsu S (unpublished work). Here PDB denotes the Protein Data Bank.

4. Comparison and discussion

The comparison was made between a 0–20 cycles SAD/SIR iteration and the combination of a 0–10 cycles SAD/SIR iteration and a 1–10 cycles MR iteration. All the iteration jobs were run automatically using the graphical user's interface of OASIS4.0[2] with default controlling parameters. SAD/SIR iteration from cycle 0 to cycle 20 was first performed. The growth of partial models from cycle to cycle is recorded as the gray curves in Figs. 1–4 for Azurin, Set7/9, TTHA1012 and Rpe respectively. The ordinates of the figures represent percentage of assigned residues, i.e. the ratio between the number of assigned residues in the built model and the total number of independent residues in the protein structure. Then,

10 cycles of MR iteration, numbered as cycles 11–20, were performed starting from the best model found in cycles 0–10 of SAD/SIR iteration. The growths of partial models are recorded as the black curves in Figs. 1–4. As is seen, in all the four figures the black curve is on average well above the gray one not only in comparison with the gray curve within the range from cycle 0 to cycle 10, but also in comparison within the range from cycle 11 to cycle 20. It is more important that for all samples the highest peak of the black curve is evidently higher than that of the gray one. This means that the combination of 11 cycles SAD/SIR iteration and 10 cycles MR iteration is obviously better than 21 cycles SAD/SIR iteration. More detailed comparison can be seen in Table 2, in which results from the best model, obtained from 0–10 cycles SAD/SIR iteration, are listed under column label I; results from the best model, obtained from 11–20 cycles SAD/SIR iteration, are listed under column label II; results from the best model, obtained within 1–10 cycles MR iteration

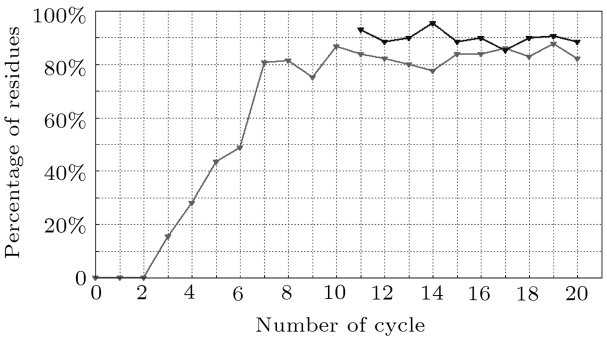


Fig. 1. Growth of structure model during the iteration for Azurin.

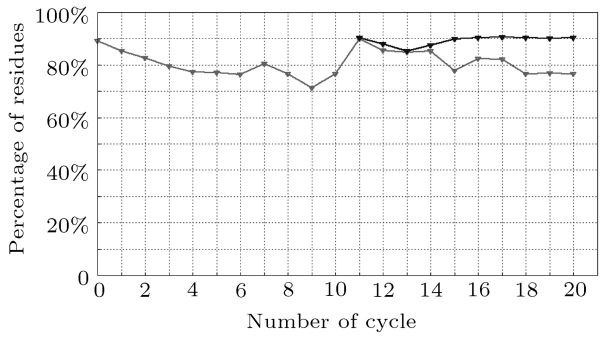


Fig. 2. Growth of structure model during the iteration for Set7/9.

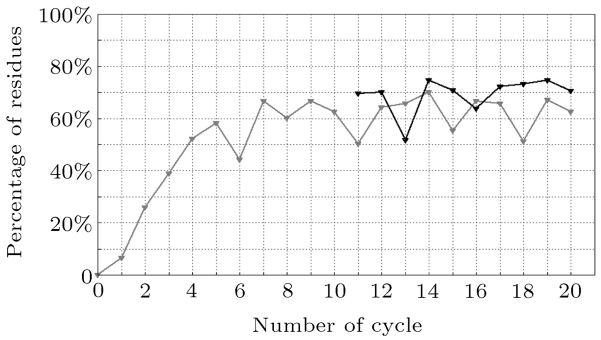


Fig. 3. Growth of structure model during the iteration for TTHA1012.

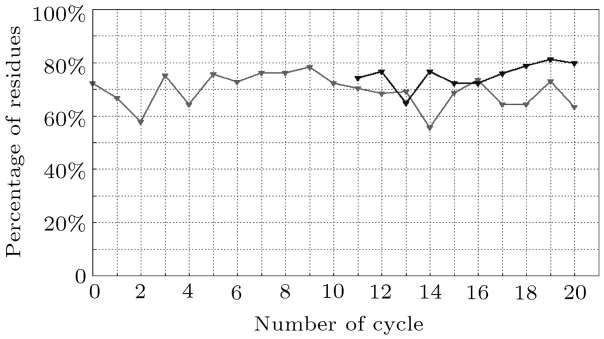


Fig. 4. Growth of structure model during the iteration for Rpe.

(corresponding to 11–20 cycles iteration of the black curves in Figs. 1–4) are listed under column label III. As is seen, the model under column label III is bigger than the larger one under column labels I and II by 9% for Azurin, 0.7% for Set7/9, 7% for TTHA1012 and 4% for Rpe. Among the four test samples, Azurin and TTHA1012 are the two most difficult cases. They gained much more from the combination of SAD/SIR and MR iterations in comparison with the other two.

Figure 5 shows what significant changes are caused by the 9% increase of the model size of Azurin. Figure 6 shows the changes of TTHA1012 caused by the 7% increase of the model size. While the sample Set7/9 has the smallest gain (0.7%) in model size, it gained the biggest decrease (more than 8 degrees) in the averaged phase error. This means that while the structure model has not increased significantly in size, the accuracy of atomic parameters (particularly in positions) has been greatly improved. As for the sample Rpe, although the 4% increase in model size is not great, it still amounts to 20 more residues assigned to the model. This is not a negligible improvement.

Table 2. Summary of test results.

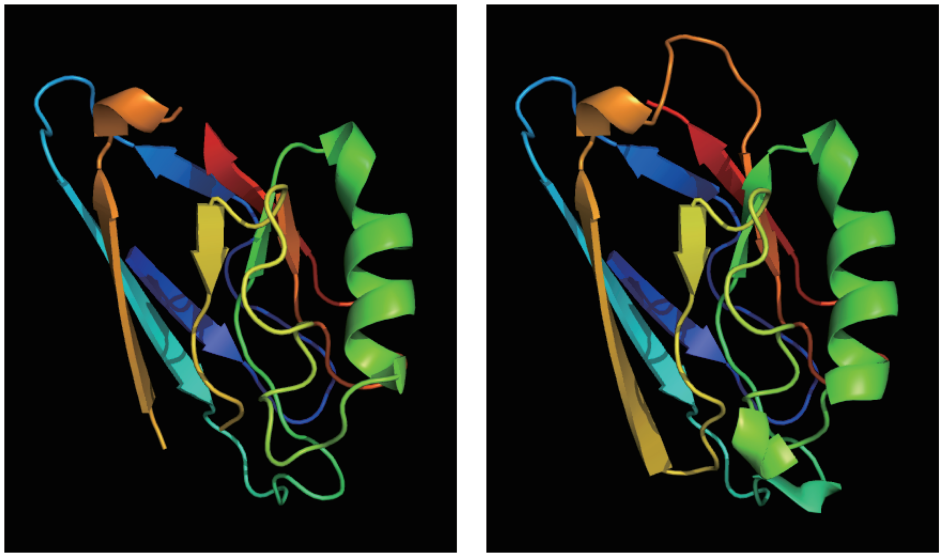
protein	number of residues assigned into the sequence			R factor/R-free factor			overall phase error of the built model/(°)		
	I	II	III	I	II	III	I	II	III
Azurin	112	113	123	0.249/0.341	0.244/0.327	0.234/0.292	29.6	29.6	28.6
Set7/9	522	527	531	0.25/0.28	0.25/0.28	0.24/0.27	40.1	38.9	30.5
TTHA1012	142	149	159	0.308/0.367	0.335/0.416	0.251/0.336	36.9	42.8	30.5
Rpe	522	490	542	0.29/0.33	0.29/0.34	0.28/0.33	39.7	40.0	39.2

I - results from the best model obtained from 0–10 cycles SAD/SIR iteration.

II - results from the best model obtained from 11–20 cycles SAD/SIR iteration.

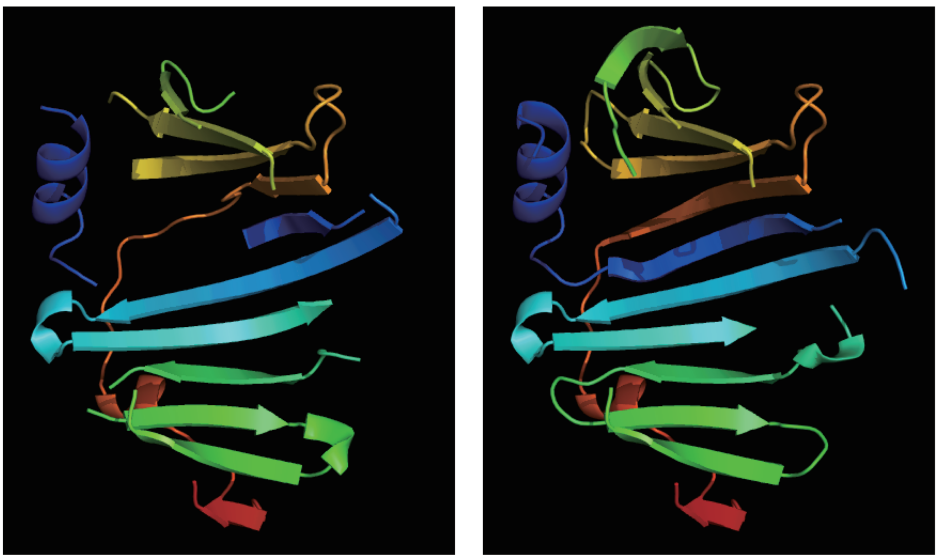
III - results from the best model obtained from 1–10 cycles MR iteration (corresponding to 11–20 cycles iterations of the black curves in Figs. 1–4) starting with the best model from 0–10 cycles SAD/SIR iteration.

Overall phase errors were calculated against the structure model in the PDB reported by the original authors.



(a) (b)

Fig. 5. Ribbon structure models of Azurin plotted by PyMOL.[19] (a) The best model from 0–20 cycles SAD iteration; (b) the best model from 1–10 cycles MR iteration based on 0–10 cycles SAD iteration.



(a) (b)

Fig. 6. Ribbon structure models of TTHA1012 plotted by PyMOL.[19] (a) The best model from cycles 0–20 cycles SAD iteration; (b) the best model from 1–10 cycles MR iteration based on 0–10 cycles SAD iteration.

5. Conclusion

For data containing SAD/SIR signals, especially when they are weak and contain large experimental errors, the combination of SAD/SIR iteration and MR iteration leads to significantly better results than that obtainable from the SAD/SIR iteration alone. This new procedure will be automated in the next version of the program OASIS.

References

[1] Wang J W, Chen J R, Gu Y X, Zheng C D and Fan H F 2004 *Acta Cryst.* D **60** 1991

[2] Zhang T, Wu L J, He Y, Wang J W, Zheng C D, Hao Q, Gu Y X and Fan H F 2009 *OASIS4.0 - a Direct-Methods Program for SAD/SIR Phasing and Reciprocal-Space Fragment Extension* Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China (The program is available free of charge for academic users at http://cryst.iphy.ac.cn)

[3] Cowtan K D and Main P 1993 *Acta Cryst.* D **49** 148

[4] Collaborative Computational Project No. 4 1994 *Acta Cryst.* D **50** 760

[5] Terwilliger T C 2000 *Acta Cryst.* D **56** 965

[6] Terwilliger T C 2003 *Acta Cryst.* D **59** 38

[7] Terwilliger T C 2003 *Acta Cryst.* D **59** 45

[8] Perrakis A, Morris R and Lamzin V S 1999 *Nature Struct. Biol.* **6** 458

[9] Vagin A A, Steiner R S, Lebedev A A, Potterton L, McNicholas S, Long F and Murshudov G N 2004 *Acta Cryst.* D **60** 2184

[10] Adams P D, Afonine P V, Bunkóczi G, Chen V B, Davis I W, Echols N, Headd J J, Hung L-W, Kapral G J, Grosse-Kunstleve R W, McCoy A J, Moriarty N W, Oeffner R, Read R J, Richardson D C, Richardson J S, Terwilliger T C and Zwart P H 2010 *Acta Cryst.* D **66** 213

[11] Yao D Q, Huang S, Wang J W, Gu Y X, Zheng C D, Fan H F, Watanabe N and Tanaka I 2006 *Acta Cryst.* D **62** 883

[12] Yao D Q, Li H, Chen Q, Gu Y X, Zheng C D, Lin Z J, Fan H F, Watanabe N and Sha B D 2008 *Chin. Phys.* B **17** 1

[13] He Y, Gu Y X, Lin Z J, Zheng C D and Fan H F 2007 *Chin. Phys.* **16** 3022

[14] He Y, Yao D Q, Gu Y X, Lin Z J, Zheng C D and Fan H F 2007 *Acta Cryst.* D **63** 793

[15] Panjikar S, Parthasarathy V, Lamzin V S, Weiss M S and Tucker P A 2009 *Acta Cryst.* D **65** 1089

[16] Dodd F, Hasnain S S, Abraham Z H, Eady R R and Smith B E 1995 *Acta Cryst.* D **51** 1052

[17] Wilson J R, Jing C, Walker P A, Martin S R, Howell S A, Blackburn G M, Gamblin S J and Xiao B 2002 *Cell* **111** 105

[18] Chang W R, Jiang T, Wan Z L, Zhang J P, Yang Z X and Liang D C 1996 *J. Mol. Biol.* **262** 721

[19] DeLano W L 2002 *The PyMOL Molecular Graphics System.* DeLano Scientific, San Carlos, CA, USA