Direct demodulation method for heavy atom position determination in protein crystallography *

ZHOU Liang(周亮) LIU Zhong-Chuan(刘忠川) LIU Peng(刘鹏) DONG Yu-Hui(董宇辉)[1)]

Beijing Synchrotron Radiation Facility, Institute of High Energy Physics,
Chinese Academy of Sciences, Beijing 100049, China

Abstract: The first step of phasing in any de novo protein structure determination using isomorphous replacement (IR) or anomalous scattering (AD) experiments is to find heavy atom positions. Traditionally, heavy atom positions can be solved by inspecting the difference Patterson maps. Due to the weak signals in isomorphous or anomalous differences and the noisy background in the Patterson map, the search for heavy atoms may become difficult. Here, the direct demodulation (DD) method is applied to the difference Patterson maps to reduce the noisy backgrounds and sharpen the signal peaks. The real space Patterson search by using these optimized maps can locate the heavy atom positions more accurately. It is anticipated that the direct demodulation method can assist in heavy atom position determination and facilitate the de novo structure determination of proteins.

Key words: direct demodulation method, heavy atom sites, difference Patterson map, protein crystallography

PACS: 61.05.C- **DOI:** 10.1088/1674-1137/37/1/018002

1 Introduction

Solution of the phase problem is central to crystallographic structure determination. In protein crystallography, the phase problem can be solved by isomorphous replacement (IR) [1, 2] or anomalous scattering (AD) [3, 4] experiments. Finding the positions of heavy atoms is a crucial step in any de novo protein structure determination using IR or AD. This step is carried out by inspecting the peaks on difference Patterson maps, which are calculated based on the derivative and native data in IR or anomalous signals in AD. In many cases this is an efficient method to locate one or a few heavy atom sites. Presently, there have already been some programs with the aim of finding heavy atom positions from difference Patterson maps. Like RSPS [5], a program for inspection and interpretation of the Patterson function in CCP4 suite [6], is mainly based on vector-search methods in the real Patterson space to locate the heavy atom sites. Another program, SOLVE/RESOLVE [7, 8], which is more widely used, is an automatic procedure from the IR/AD data to the electron density maps and model-building. SOLVE mainly uses difference Patterson maps to determine the heavy atom positions in IR or AD data sets. The only input for the program SOLVE can be the diffraction data; the positions of the heavy atoms can be identified and then the initial phases can be calculated from the solutions of heavy atoms [7]. RESOLVE is used to apply density modifications to improve the phases and model-building [8].

However, with the increasing complexity of the structures or the low data quality collected by X-ray detectors, the difference Patterson maps calculated from the diffraction data usually have higher backgrounds or the crowding of numerous signal peaks. In such cases, interpreting the difference Patterson maps may become more difficult, since we have to identify great numbers of weak peaks from strong backgrounds. In the fields of image restoration and computerized tomography, a method has been successfully used to extract the signals from incomplete and noisy data. This method is the direct demodulation (DD) method [9, 10]. It uses an iterative algorithm to solve the modulation equations under physical constraints and is proved to be a powerful technique for dealing with the inverse problem [11, 12]. It is possible to use the direct demodulation method to optimize the difference Patterson maps to reduce the background noises and sharpen the signal peaks in the maps, so as to locate the heavy atom positions correctly and accurately for the phasing in the structural determination. In this paper, we applied the direct demodulation method to the difference Patterson map, and then successfully

Received 9 March 2012, Revised 7 May 2012

* Supported by National Natural Science Foundation of China (10979005) and National Basic Research Program of China (2009CB918600)

1) Correspondence E-mail: dongyh@ihep.ac.cn

locate the heavy atom sites. The results of phasing are better than those by using conventional programs such as SOLVE/RESOLVE.

2 Applying the direct demodulation method to a difference Patterson map

In the field of high energy astronomy image restoration, the real object distribution between the observed data in one-dimensional space can be mathematically described by [9]:

$$\sum_{i=1}^{N} P_{i,i'} f(i) = d(i') i' = 1,2,\cdots,M, \tag{1}$$

or in matrix form

$$Pf = d, \tag{2}$$

where $f(i)$, $i=1,\cdots,N$ is the real object distribution at a point i and $d(i')$ $i'=1,\cdots,M$ represent the observed data at an observation point i'. $P_{i,i'}$ is the point-spread function of the detector system, which can be regarded as the response of the detector to a point i of object space during the observation point i'.

The most straightforward way to evaluate an object from observed data should directly solve the modulation Eq. (1). But errors always exist in modulation equations, such as statistical fluctuation and noises in observation data, etc. The mathematical solutions will seriously deviate from the true object distribution and violently oscillate. Li and Wu suggested retrieving the object from observed data by solving the modulation Eq. (1) iteratively under physical constraints. This is the direct demodulation method [9]. If the point-spread function for any object point is relatively concentrated round the corresponding observation point, a normal iteration method, e.g., the Gauss-Seidel algorithm can be used to solve the modulation Eq. (1). The approximate solution for l-th iteration can be calculated by using the following formulas [9]:

$$f(i)^{(l)} = \frac{\alpha}{P'_{ii}} \left(d_i - \sum_{j=1}^{i-1} P_{ij} f(j)^{(l)} - \sum_{j=i+1}^{N} P_{ij} f(j)^{(l-1)} \right) + (1-\alpha) f(j)^{(l-1)}, \tag{3}$$

where the relaxation factor $0<\alpha<1$. It has demonstrated that setting a reasonable physical constraint in the iterative process can effectively strengthen the convergence and depress the influence of noise. Therefore, we can set up a lower bound b and upper bound u, for any approximate solution $f^{(l)}(i)$:

if $f^{(l)}(i)<b(i)$, let $f^{(l)}(i)=b(i)$;
if $f^{(l)}(i)>u(i)$, let $f^{(l)}(i)=u(i)$.

The difference Patterson map in protein crystallography indicates the vectors between the heavy atoms inside the proteins. As we know, the peaks in the difference Patterson map are the convolution of the electron densities around the heavy atoms which define the vectors. It is reasonable that the electron densities around the heavy atoms should follow the Gaussian distribution, therefore the peaks in a difference Patterson map can also be processed as a series of discrete peaks spread by a Gaussian point-spread function. On this premise, the peaks $d(u', v', w')$ observed in the difference Patterson map can be regarded as the convolution of signal peak $f(u,v,w)$ by a Gaussian distribution function $p_{uu',vv',ww'}$. Expanding Eq. (1) to the 3-dimensional Patterson space, the modulation equation can be derived as:

$$\sum_{u}\sum_{v}\sum_{w} p_{uu',vv',ww'} f(u,v,w) = d(u',v',w'). \tag{4}$$

Because of the convolution of the Gaussian distribution function, the peaks in the Patterson map have a Gaussian broadening. When the structures of proteins become large or the symmetries of crystals are high, the peak broadening may cause the crowding of peaks and possible peak overlaps. Also, there is some noisy background associated with the observed peak $d(u', v', w')$ due to the large numbers of vectors between the light atoms (most atoms of proteins are C, N, O and S) inside proteins and the inevitable experimental error in the diffraction data. In this case, interpreting the difference Patterson maps may become difficult. Under this situation, the direct demodulation method can be applied to the difference Patterson map.

The iteration method based on the Gauss-Seidel method can be used to solve the modulation Eq. (4) in a 3-dimensional Patterson space, which is described as:

$$\begin{aligned} f^{(l)}(u,v,w) = {} & \frac{\alpha}{p_{uu,vv,ww}} d(u,v,w) + (1-\alpha) f^{(l-1)}(u,v,w) \\ & - \frac{\alpha}{p_{uu,vv,ww}} \left(\sum_{w'=w-m}^{w-1} \sum_{v'=v-m}^{v+m} \sum_{u'=u-m}^{u+m} p_{uu',vv',ww'} f^{(l)}(u',v',w') \right. \\ & \left. + \sum_{w'=w+1}^{w+m} \sum_{v'=v-m}^{v+m} \sum_{u'=u-m}^{u+m} p_{uu',vv',ww'} f^{(l-1)}(u',v',w') \right. \end{aligned}$$

$$+\sum_{v'=v-m}^{v-1}\sum_{u'=u-m}^{u+m} p_{uu',vv',ww} f^{(l)}(u',v',w)+\sum_{v'=v+1}^{v+m}\sum_{u'=u-m}^{u+m} p_{uu',vv',ww} f^{(l)}(u',v',w)$$

$$+\sum_{u'=u-m}^{u-1} p_{uu',vv,ww} f^{(l)}(u',v,w)+\sum_{u'=u+1}^{u+m} p_{uu',vv,ww} f^{(l-1)}(u',v,w)\Bigg), \qquad (5)$$

where the relaxation factor $0<\alpha<1$, and m represents the Gaussian broadening in each peak. Concerning the convergence of iteration process, we monitor the maximum deviation of the neighboring two solutions ε during the iterations, described as Eq. (6). If ε is less than an acceptable value, the solution can be considered as a convergent result. The optimal value of α is a heuristic number, and the value of ε represents the precision of the solution. In our test, the final solution has no obvious difference relevant to the selection of α. So α is set to 0.9, ε is set to 0.1, which can maintain the speed of convergence and a high precision of solution. After a few iterations (usually less than 100 cycles), the convergent result is usually reached.

$$\varepsilon = \max_{0<u'<u,0<v'<v,0<w'<w}(|f^{(l)}(u',v',w')$$

$$-f^{(l-1)}(u',v',w')|). \qquad (6)$$

By employing the DD method to the difference Patterson map, we have downloaded 2 diffraction data by AD from 2 known protein structures in the Protein Data Bank (PDB bank). The original difference Patterson map can be easily calculated from the diffraction data. Summarized in Table 1, the crystal structure of a putative aminotransferase from Silicibacter pomeroyi (PDB ID: 3H14) was determined at 1.9 Å resolution, which has 11 Se in the asymmetric unit. And the crystal structure of a putative type 11 methyltransferase from Sulfolobus solfataricus (PDB ID: 3I9F) was determined at 2.5 Å resolution, which has six Zn in the asymmetric unit. The Bijvoet ratios are 2.28% and 4.65% respectively, which indicates an overall good signal from the heavy atom in the difference Patterson map.

The feasibility of the theory applied to the difference Patterson map can be depicted in the case of a protein structure whose PDB ID is 3H14. Fig. 1(a) shows a cross section of the original difference Patterson map calculated from the diffraction data. We can see there are some peaks together with a noisy background in the Patterson space. The noisy background has a level from −1.5 to 1, due to the large number of vectors between the light atoms and the experiment error in the diffraction data. Then the DD method is applied to this difference Patterson map. The non-negative constraints are used in the iterative process to solve Eq. (5), e.g., if $f^{(l)}(u,v,w)<0$, then let $f^{(l)}(u,v,w)=0$. After the optimization, a clearer and sharpener map was obtained. As shown in Fig. 1(b), only a few peaks are left in the map, it is clearly seen that the peak height is higher than the corresponding peak in the original map. The main effect of the DD method is to reduce the noisy backgrounds and sharpen the level of signal peaks in the original difference Patterson map. It should be noted that there are still some fake peaks left in this optimized map. These fake peaks could cause fake heavy atom positions, but they can be effectively rejected by RSPS during the vector-search process [5].

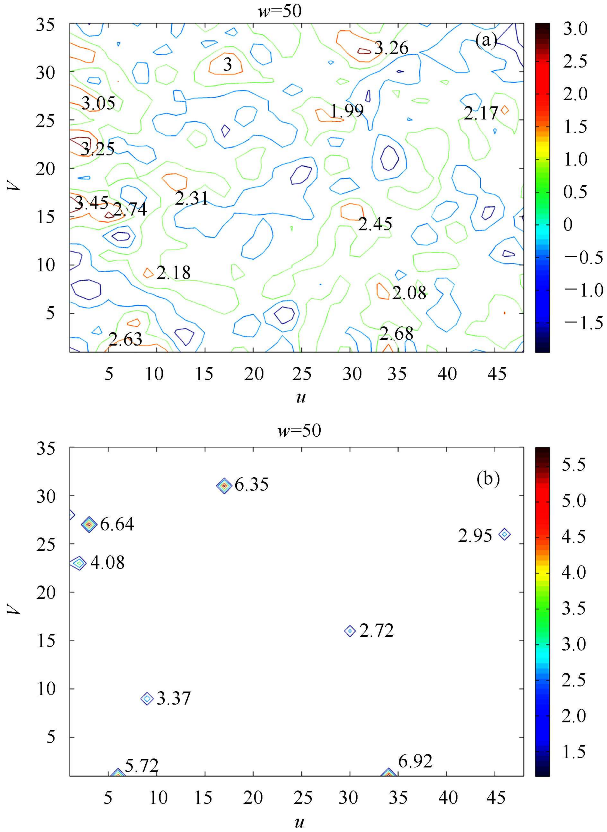

Fig. 1. (color online) Comparison between the original difference Patterson map (a) and the DD optimized map (b) on a cross section w=50 in the case of 3H14. Only the peak height is labeled.

Table 1. Three cases of protein downloaded from the PDB.

ID	space group	resolution/Å	number of residues	heavy atom	wavelength/Å	Bijvoet ratio ($\langle\|\Delta F\|\rangle/\langle F\rangle$)(%)
3H14	C222$_1$	1.9	391	Se(11)	0.97929	2.28
3I9F	P2$_1$3	2.5	170	Zn(6)	0.97958	4.65

$*$ Bijvoet ratio: $\langle|\Delta F|\rangle/\langle F\rangle\approx(N_{\rm A}/N_{\rm T})^{1/2}(2f''/Z_{\rm eff})$. $N_{\rm A}$ is the number of anomalous scatters, $N_{\rm T}$ is the total number of atoms in the structure and $Z_{\rm eff}$ is the normal scattering power for all atoms

3 Results and discussion

We use three different procedures to find the heavy atom location in the two cases for comparison:

(1) RSPS [6]: The original difference Patterson maps calculated from the diffraction data were inputted into the RSPS program to find heavy atom sites.

(2) DD-RSPS: The original difference Patterson map was first optimized by the DD method, and then inputted into the program RSPS.

(3) SOLVE [7]: The diffraction data downloaded from the PDB were directly inputted into the SOLVE program for automatic heavy atom sites determination.

The results of three procedures are listed in Table 2. Due to the noisy backgrounds and the weak peaks in the original difference Patterson map, it fails to locate the heavy atom positions with this map by RSPS. In contrast, the RSPS program functions well with the optimized map by using the DD-RSPS procedure in the 2 cases. Compared with the solution from SOLVE, the DD-RSPS procedure can find nearly the same numbers of heavy atom as SOLVE does. So the optimized map, which has lower noisy backgrounds and sharpener signal peaks, is more suitable and effective for performing the real space Patterson search than the original Patterson map.

In our test，the DD-RSPS procedure can get a better solution of heavy atom position determination. For the case of 3I9F, six Zn atoms were determined by both DD-RSPS and SOLVE. As listed in the last column of Table 3, the heavy atom substructures of 3I9F all have a high occupancy, which indicates the good signal strength of each heavy atom. This high quality signal is favorable for locating the heavy atom positions. The sites of Zn atoms are completely the same when determined by the 2 different procedures (see Table 3). For the case of 3H14, a full set of Se atoms can be located by DD-RSPS, but only 10 of 11 Se atoms can be found by SOLVE automatically. Illustrated in Table 4, it is seen that the 11$^{\rm st}$ Se atom has a lower occupancy of 0.3844 than the other heavy atom. That is to say, the signal strength from this atom is lower than the other heavy atom signal. This leads to the failure to locate the 11$^{\rm st}$ Se atom by SOLVE. After optimization by the DD method, the map has lower noisy backgrounds and sharpener signal peaks compared with the original map (see Fig. 1(b)). With this optimized map, we can effectively perform the real space Patterson search and successfully locate the 11$^{\rm st}$ Se atom by RSPS.

Table 2. Results of the RSPS, DD-RSPS and SOLVE procedure to find the position of heavy atom position.

ID	numbers of heavy atoms found by		
	RSPS	DD-RSPS	SOLVE
3I9F	0	6	6
3H14	0	11	10

Then, the 2 sets of heavy atom parameters determined by the DD-RSPS/SOLVE of each case are used for structural determination. The heavy atom parameters and diffraction data were inputted into the SOLVE (phasing only)-RESOLVE procedure for phasing. RESOLVE was used for density modification and model-building in all cases. The resultant overall-averaged phase errors are listed in Table 5.

It is seen that in the case of 3I9F, the overall-averaged phase errors from different procedures are nearly the same. That is because the two sets of heavy atom parameters determined by DD-RSPS and SOLVE are completely the same in this case. Moreover, the DD-RSPS-RESOLVE procedure could lead to a result better than that of SOLVE-RESOLVE. For the case of 3H14, the phasing result has been improved through DD-RSPS-RESOLVE procedure by more than 2.1 degrees.

Table 3. Comparison between the solutions of the DD-RSPS and SOLVE procedures to find the heavy atom position on a fractional coordinate in the case of 3I9F. The position of the 1$^{\rm st}$ Zn determined by DD-RSPS is equivalent to the position determined by SOLVE because of the symmetry P2$_1$3.

Zn	DD-RSPS			SOLVE			q
	X	Y	Z	X	Y	Z	
1	0.0930	0.0030	0.0938	0.0938	0.0930	0.0030	1.0630
2	0.3221	0.2666	0.1954	0.3222	0.2666	0.1954	1.1504
3	0.0384	0.1645	0.1236	0.0383	0.1645	0.1236	0.8747
4	0.7068	0.7625	0.1344	0.7067	0.7624	0.1344	1.0300
5	0.1573	0.9414	0.0838	0.1572	0.9414	0.0838	0.9078
6	0.2519	0.4372	0.2398	0.2519	0.2400	0.2400	0.8633

$*X$, Y, Z: fractional coordinates; q: occupancy

Table 4. Comparison between the solutions of the DD-RSPS and SOLVE procedures to find the anomalous scatters position on a fractional coordinate in the case of 3H14.

Se	DD-RSPS			SOLVE			q
	X	Y	Z	X	Y	Z	
1	0.5572	0.2332	0.1853	0.5573	0.2332	0.1853	1.0514
2	0.2397	0.2930	0.0678	0.2397	0.2930	0.0678	1.1959
3	0.5679	0.3647	0.0570	0.5678	0.3647	0.0570	0.7741
4	0.3002	0.2949	0.0445	0.3002	0.2950	0.0445	0.7620
5	0.6584	0.4222	0.2053	0.6584	0.4222	0.2053	0.8977
6	0.4185	0.4744	0.2337	0.4185	0.4743	0.2337	0.7148
7	0.5551	0.4786	0.1533	0.5550	0.4786	0.1534	0.5196
8	0.1109	0.3694	0.0216	0.1108	0.3694	0.0216	0.8171
9	0.6726	0.4648	0.0914	0.6725	0.4648	0.0914	0.7879
10	0.2363	0.3551	0.1619	0.2363	0.3551	0.1619	0.6882
11	0.2577	0.3882	0.1832				0.3844

*X, Y, Z: fractional coordinates; q: occupancy

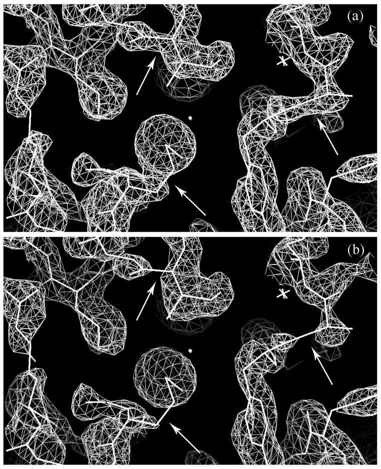

Fig. 2. 1.9 Å electron-density maps of the case 3H14 contoured at 1σ by (a) DD-RSPS-RESOLVE (b) SOLVE-RESOLVE procedure. The known structure (shown in stick mode) is superimposed. As is seen by the arrow, the electron density map derived from the DD-RSPS-RESOLVE procedure provides a better connectivity of the main-chain density to the known structure than that of SOLVE-RESOLVE.

Although the improvement in this case is only about 2.1 degrees, the effect on the corresponding Fourier electron-density map is evident. Fig. 2 shows the electron-density map output by the two different procedures, respectively, the known structure is shown in stick model. In comparison with the known stick model superimposed, it is seen that the electron-density map from the DD-RSPS-RESOLVE procedure (see Fig. 2(a)) provides much more structural information than that of SOLVE/RESOLVE (see Fig. 2(b)) So the electron-density map derived from DD-RSPS-RESOLVE procedure is much easier to interpret than that from SOLVE-RESOLVE.

Table 5. Overall-averaged phase errors in degrees of different phasing procedures in the 2 cases.

procedure	ID	
	3I9F	3H14
SOLVE-RESOLVE	62.0	64.8
DD-RSPS-RESOLVE	62.1	62.7

*The overall-averaged phase errors are calculated between the phasing result by SOLVE and the Fourier transformation of the known structural model

4 Conclusion

The present test shows that the direct demodulation method can optimize the difference Patterson map. With this optimized map, we can locate the positions of heavy atoms accurately by using a real space Patterson search method. The phasing result is better than conventional programs such as SOLVE/RESOLVE. It is anticipated that the direct demodulation method can play an assistant role in the heavy atom position determination and facilitate the de novo structure determination of proteins.

References

1 Perutz M F. Acta Cryst, 1956, **9**: 867

2 Kendrew J C, Bodo G, Dintzis H M, Parrish R G, Wyckoff H, Phillips D C. Nature (London), 1958, **181**: 662

3 Hendrickson W A, Smith J L, Phizackerley R P, Merritt E A. Proteins, 1988, **4**: 77

4 Murthy H M, Hendrickson W A, Orme-Johnson W H, Merritt E A, Phizackerley R P. J. Biol. Chem, 1988, **263**: 430

5 Stefan D. Knight. Acta Cryst. D, 2000, **56**: 42

6 Collaborative Computational Project, Number 4. Acta Cryst. D, 1994, **50**: 760

7 Terwilliger T C. Acta Cryst. D, 1999, **55**: 1863

8 Terwilliger T C. Acta Cryst. D, 2000, **56**: 965

9 LI T P, WU M. Astrophys. Space Sci., 1993, **206**: 91

10 LI T P, WU M. Astrophys. Space Sci., 1994, **15**: 213

11 ZHANG S, LI T P, WU M. Acta Astrophysica Sinica, 1997, **17**(3): 263

12 CHENG Y, SONG L, LI T P et al. Acta Astronomica Sinica, 2000, **41**(2): 214