Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 31 August 2011
Accepted 14 December 2011

Phase retrieval in protein crystallography

Zhong-Chuan Liu, Rui Xu and Yu-Hui Dong*

Beijing Synchrotron Radiation Facility, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, People's Republic of China. Correspondence e-mail: dongyh@ihep.ac.cn

Solution of the phase problem is central to crystallographic structure determination. An oversampling method is proposed, based on the hybrid input–output algorithm (HIO) [Fienup (1982). *Appl. Opt.* **21**, 2758–2769], to retrieve the phases of reflections in crystallography. This method can extend low-resolution structures to higher resolution for structure determination of proteins without additional sample preparation. The method requires an envelope of the protein which divides a unit cell into the density region where the proteins are located and the non-density region occupied by solvents. After a few hundred to a few thousand iterations, the correct phases and density maps are recovered. The method has been used successfully in several cases to retrieve the phases from the experimental X-ray diffraction data and the envelopes of proteins constructed from structure files downloaded from the Protein Data Bank. It is hoped that this method will greatly facilitate the *ab initio* structure determination of proteins.

1. Introduction

In the standard procedures of protein crystallography, one of the most difficult problems is to obtain the phases from the available moduli of the structure factors. In diffraction experiments, the reflection intensities $I(h, k, l)$ can be measured on the detector. The amplitudes of the structure factors, $|F(h, k, l)|$, are proportional to the square roots of the intensities. However, the phases are lost in the experiments. To reconstruct the electron density of the object, the phase information must be known. This constitutes the well known phase problem.

A number of phasing strategies have been developed that in many situations perform very well. The first method to determine the phases for protein crystals was multiple isomorphous replacement (MIR) introduced by Perutz and Kendrew (Perutz, 1956; Kendrew *et al.*, 1958). Furthermore, if some similar protein structures are known in advance, molecular replacement may provide the solutions for the phase problem (Rossmann & Blow, 1962). With the advent of synchrotron radiation and the development of molecular biology techniques, a new method of multiple-wavelength anomalous dispersion (MAD) has been introduced (Hendrickson *et al.*, 1988; Murthy *et al.*, 1988). Additionally, even single isomorphous replacement (SIR) or single-wavelength anomalous dispersion (SAD) can solve the phases in protein crystallography (Wang, 1985) or with the aid of direct methods (Ealick, 1997; Weeks & Miller, 1999). However, isomorphous replacement methods (MIR and SIR) rely on the stringent preparation of heavy-atom derivative crystals. Likewise, anomalous scattering techniques (MAD and SAD) require the existence of heavy atoms in proteins, and sometimes Se atoms need to be introduced during the production of proteins. Nonetheless, to obtain the phases in the *de novo* structure determination of proteins without additional sample preparation always attracts the interest of crystallographers.

The difficulties in *de novo* structure determination of proteins are due to the intrinsic insufficiency of the diffraction data observed in experiments. One can obtain diffraction intensities $I(h, k, l)$ from a crystal with a grid of $N_1 \times N_2 \times N_3$ points along the h, k and l directions in reciprocal space; therefore, the electronic densities can be calculated on the $N_1 \times N_2 \times N_3$ points along the x, y and z directions in real space, if the phases of the diffraction spots can be known. Owing to the nature of the crystal, these $N_1 \times N_2 \times N_3$ diffraction peaks appearing in three dimensions must satisfy Laue's law. Also the values of the diffraction intensities are constrained by Friedel's law of diffraction, where $I(-h, -k, -l) = I(h, k, l)$ and $F(-h, -k, -l) = F(h, k, l)^*$. Generally speaking, we have to determine the values of the electron densities in the grid of $N_1 \times N_2 \times N_3$ points, but only $N_1 \times N_2 \times N_3/2$ independent intensities can be measured in experiments. The lost information is the phases of the diffraction spots, while Friedel's law gives the constraint on the phases as $\varphi(h, k, l) = 2\pi - \varphi(-h, -k, -l), 0 \leq \varphi \leq 2\pi$. In protein crystallography, the well known process of isomorphous replacement (IR), and also anomalous scattering (or AD, anomalous dispersion), provides an opportunity to measure more data. For example, the anomalous scattering breaks Friedel's law, where $|F(-h, -k, -l)|$ is no longer equal to $|F(h, k, l)|$. Hence we can measure $N_1 \times N_2 \times N_3$ independent diffraction intensities

instead of $N_1 \times N_2 \times N_3/2$. Likewise, isomorphous replacement also implements the measurement of another set of $N_1 \times N_2 \times N_3/2$ independent data points. That is sufficient to explain why IR or AD methods are applicable for solving the phase problem in protein crystallography.

Several previous works concerning the *ab initio* phasing method based on one set of non-anomalous data alone have been widely discussed. Lunin *et al.* (2000) reported the method of solving the phases of low-resolution data and could further extend the low-resolution phases to high resolution when very high non-crystalline symmetry is present. Hao (2001) applies a six-dimensional search to locate the envelope determined by SAXS (small-angle X-ray scattering), a method developed by Svergun & Stuhrmann (1991), to yield the low-resolution structure of protein crystals. Unfortunately, the low-resolution phases, which correspond with these low-resolution structures, are difficult to expand to high resolution. Miao *et al.* (2000) introduce an oversampling phasing method for high-resolution three-dimensional structure determination of complex and non-crystalline biological specimens. By employing an iterative algorithm, the phase information from the oversampled diffraction pattern of a micrometre-sized test specimen has been successfully retrieved. Also, for crystals Miao & Sayre (2000) showed there are possibilities for oversampling. In principle, it is very likely that in protein crystals proteins only occupy half of the volume inside the crystal cells since the solvent contents are usually higher than 50%. In such cases, we can measure $N_1 \times N_2 \times N_3/2$ independent diffraction intensities and only less than $N_1 \times N_2 \times N_3/2$ unknown values of electron densities need to be determined; the numbers of measured independent diffraction intensities can be more than the numbers of unknown electron densities where the to-be-determined structure in the object domain will be surrounded by mathematically zero solvent. Therefore, it is possible to use the oversampling method for retrieving the phase information.

Here we propose the oversampling method in protein crystallography to retrieve the phases of diffraction. Based on the hybrid input–output algorithm (HIO) (Fienup, 1982), the phases of diffraction peaks of the crystals whose solvent contents are higher than 50% can be retrieved. The algorithm is not sensitive to the errors in the diffraction intensities, while some processes are necessary for correct retrieval. This algorithm is successfully applied to diffraction data sets of high-solvent-content crystals and the phases can be obtained.

2. Method

As Hao (2001) has pointed out, after the envelopes of proteins have been determined by SAXS, both the locations and orientations of the proteins inside the crystal cells can be defined. Therefore, both the density regions D where the proteins are located and also the no-density regions occupied by solvents in the cells are known. On the other hand, the reflections of the crystals give the amplitudes of the structure factors, $|F^{\rm exp}(h, k, l)|$, as the reciprocal constraints. Here no anomalous scattering happens and Friedel's law constrains the structure factors by $|F^{\rm exp}(h, k, l)| = |F^{\rm exp}(-h, -k, -l)|$.

In such a case, the phases of the reflections can be retrieved by the following algorithm.

(i) The initial electron densities in the crystal cells can be set as

$$\rho_0(x, y, z) = \begin{cases} 1, [(x, y, z) \in D] \\ 0, [(x, y, z) \notin D] \end{cases}. \quad (1)$$

Assuming the electron densities are divided into a grid of $N_1 \times N_2 \times N_3$ points, the Fourier transformation of $\rho_0(x, y, z)$ gives the structure factors $|F^{\rm calc}(h, k, l)|\exp[i\varphi(h, k, l)]$.

(ii) Replacing $|F^{\rm calc}(h, k, l)|$ by $|F^{\rm exp}(h, k, l)|$, while keeping the values of the newest phases $\varphi(h, k, l)$, a new set of structure factors $|F^{\rm exp}(h, k, l)|\exp[i\varphi(h, k, l)]$ is constructed. After applying the inverse Fourier transformation, we get a new density $\rho_1'(x, y, z)$.

(iii) Modify $\rho_1'(x, y, z)$ according to equation (2), which pushes the grid points outside the support D close to zero to fulfil the real-space constraints based on HIO (Fienup, 1982):

$$\rho_m(x, y, z) = \begin{cases} \rho_{m-1}(x, y, z), [(x, y, z) \in D] \\ \rho_{m-1}(x, y, z) - \varepsilon\rho_m'(x, y, z), [(x, y, z) \notin D] \end{cases}. \quad (2)$$

(iv) Apply the Fourier transformation to obtain a new set of $|F^{\rm calc}(h, k, l)|$ and $\varphi(h, k, l)$ as the second step of input for the successive iteration.

(v) After a few hundreds to thousands of iterations, convergences are usually reached. According to equation (3) instead of equation (2), modify $\rho(x, y, z)$ in the iterations a few times in order to push the grid points outside the support D to zero, like solvent flattening (Wang, 1985):

$$\rho_m(x, y, z) = \begin{cases} \rho_{m-1}(x, y, z), [(x, y, z) \in D] \\ 0, [(x, y, z) \notin D] \end{cases}. \quad (3)$$

Concerning the convergence of the phasing algorithm, the agreement between the true and estimated structures should be monitored during the iterations. Here we utilize the cross-correlation coefficient (CC) and average error in the phase angle $\Delta\varphi$ defined as follows to trace the process:

$$\mathrm{CC} = \frac{\sum_{hkl} \left|F^{\rm exp}(h, k, l)\right| \left|F^{\rm calc}(h, k, l)\right| \cos\left[\varphi^{\rm exp}(h, k, l) - \varphi^{\rm calc}(h, k, l)\right]}{\left[\sum_{hkl} \left|F^{\rm exp}(h, k, l)\right|^2 \sum_{hkl} \left|F^{\rm calc}(h, k, l)\right|^2\right]^{1/2}} \quad (4)$$

$$\Delta\varphi = \frac{\sum_{hkl} \arccos\left\{\cos[\varphi^{\rm exp}(h, k, l) - \varphi^{\rm calc}(h, k, l)]\right\}}{\sum_{hkl}}. \quad (5)$$

The algorithm iterates between reciprocal and real space. In reciprocal space, the correct Fourier moduli are restricted by the amplitudes of the structure factors; in real space, the electron densities outside the support region D are gradually pushed close to zero. Since the phasing algorithm alternately implemented the reciprocal-space and the real-space constraints, the correct phase set can be retrieved after a few

hundred to a few thousand iterations. Here HIO (Fienup, 1982) is proposed and the procedure is similar to solvent flipping (Abrahams, 1997). The optimal value of ε is a heuristic number between 0.5 and 1, which controls the speed of convergence. Usually ε is set to 0.9.

3. Verification of the phasing algorithm and simulation results

3.1. The feasibility of the phasing algorithm

By employing the algorithm mentioned above, we have performed computer modelling to reconstruct the electron densities in crystal cells. A number of residues were cut from a PDB (Protein Data Bank) file (1n0h), which were confined inside a finite spherical region with a radius of 20 Å. Then a translation along the z axis was applied, giving the molecule a translational symmetry. We utilized this structure to construct several crystal cells with different cell parameters and with distinct solvent contents. The constructed PDB files were used to calculate the diffraction intensities for the subsequent reconstructions of these models, where the structure factors were firstly obtained by the program *SFALL* of the CCP4 library (Collaborative Computational Project, Number 4, 1994) at 2.00 Å resolution. The first case, named P41-100, has a very high solvent content of 70.34% with a crystal cell of 100 × 100 × 105 Å. Then we reduced the dimensions of the crystal cells with 5 Å step length, in order to arrive at distinct solvent contents. The space group of all examples is $P4_1$. The final example P41-80 has a crystal cell of 80 × 80 × 85 Å and has a low solvent content of 42.99%. Details of the different examples are given in Table 1.

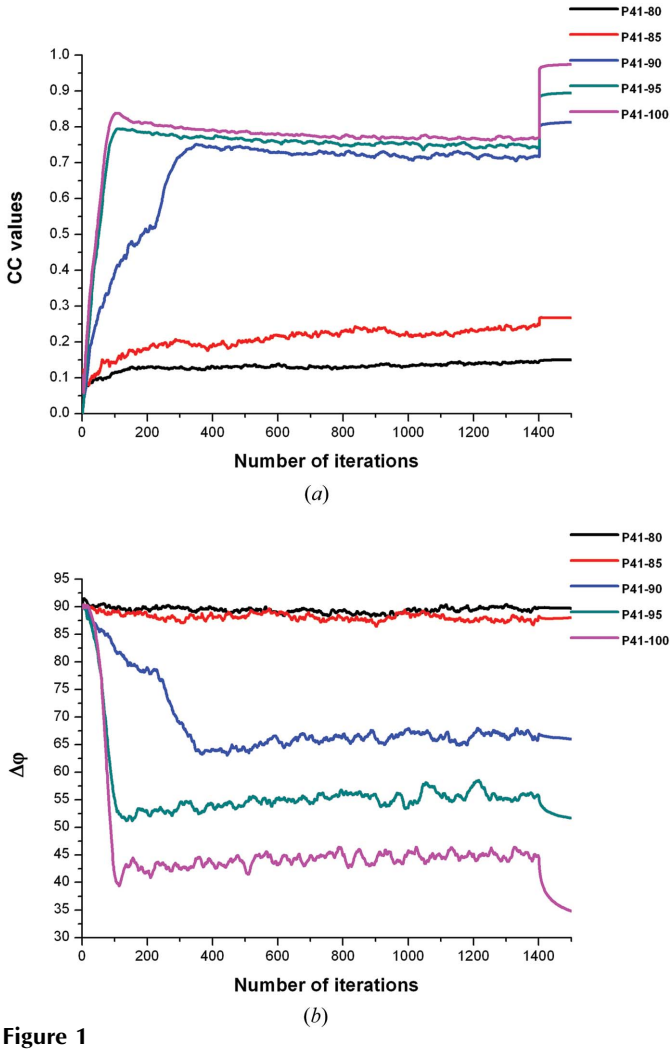


Figure 1
The convergent processes of P41-100, 95, 90, 85 and 80 after 1400 iterations of the phasing algorithm and the final stage of 100 cycles of solvent flattening without histogram matching. (*a*) CC values plotted *versus* the number of iterations, (*b*) average errors in phase angles $\Delta\varphi$ plotted *versus* the number of iterations. P41-90, 95, 100 in blue, dark cyan, magenta, respectively, all have convergent results. P41-80 in black and P41-85 in red have no obvious convergent process.

Table 1
The results of different models in testing the algorithm of phase retrieval.

HM: histogram matching.

	Dimensions of the cells (Å)	Solvent content (%)	CC		$\Delta\varphi$ (°)	
			No HM	HM	No HM	HM
P41-100	100 × 100 × 105	70.34	0.97	0.95	34.8	45.3
P41-95	95 × 95 × 100	65.75	0.89	0.87	51.6	53.8
P41-90	90 × 90 × 95	59.61	0.81	0.82	66.0	60.4
P41-85	85 × 85 × 90	52.25	0.27	0.87	87.9	57.0
P41-80	80 × 80 × 85	42.99	0.15	0.07	89.7	89.2

Fig. 1 shows the convergent process of five cases after 1400 iterations of the phasing algorithm and the final stage of 100 cycles of solvent flattening. For the cases P41-100, 95 and 90, that have a solvent content of 70.34, 65.74 and 59.61%, respectively, the CC (cross-correlation coefficient) values converged after sufficient iterations to a final CC of 0.97, 0.89 and 0.81, respectively. In the meantime the average errors in the phase angles $\Delta\varphi$ (between the recovered phases and the theoretical phases computed from the constructed PDB file by the program *SFALL*) converged to 34.8, 51.6 and 66.0°, respectively. The electron-density map of case P41-100 evolves from a homogeneous sphere to the correct densities as is illustrated in Fig. 2. It is easy to trace the polypeptide chain from the 2.00 Å density map for structure determination. On the contrary, the cases of P41-85, 80 that have lower solvent contents of 52.25 and 42.99%, respectively, have no conver-

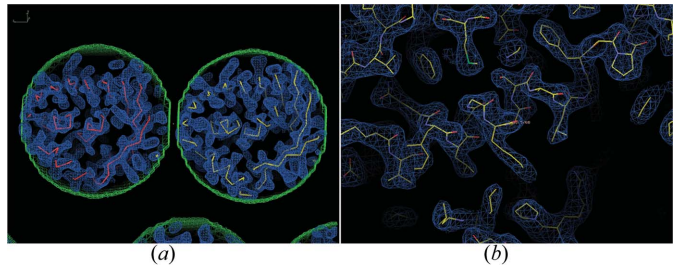


Figure 2
(*a*) The 2.0 Å electron-density map (in blue) of P41-100 after convergence compared with the α-carbon tracing of the constructed model superimposed and the envelope (in green). (*b*) A small section of the map with the final structure superimposed.

gence results, as shown in Fig. 1 by the red and black lines. Concerning the applicability of the phasing algorithm, the oversampling condition should be satisfied. This means in principle that the solvent contents in the crystal cells should be at least higher than 50%, and the higher the better. Therefore, this algorithm did not take effect for the cases of P41-85, 80, because of the restraint of high solvent contents.

3.2. Applying histogram matching

In the last section, we showed that the phasing algorithm can retrieve the phase directly from the diffraction data and the envelope of the protein. In 1988 and in subsequent works, Lunin pointed out (Lunin, 1988; Lunin & Skovoroda, 1991; Lunin *et al.*, 1990; Lunin & Vernoslova, 1991) that histogram matching could be useful for improving electron-density maps. Zhang & Main (1990*a*) presented it in a simplified form and incorporated it in the program *SQUASH* (Zhang & Main, 1990*b*); another example is *DM* in the CCP4 library (Collaborative Computational Project, Number 4, 1994; Cowtan & Main, 1996). We can also incorporate histogram matching by using the subroutine of *DM* as another real-space constraint in the iterations.

After the modification of densities according to equation (2), histogram matching is then applied to the densities. We have performed this improved method on the five cases mentioned above. Fig. 3 shows the convergent process of the five cases after 1400 iterations of the phasing algorithm and the final stage of 100 cycles of solvent flattening with histogram matching. For the cases of P41-100, 95, 90, we succeeded in retrieving the phases, and the results did not differ much compared with those obtained without histogram matching. However, as shown in Fig. 4, for the case of P41-85 which has a solvent content of 52.25% there is a clear difference in the results obtained with histogram matching (in red) and without

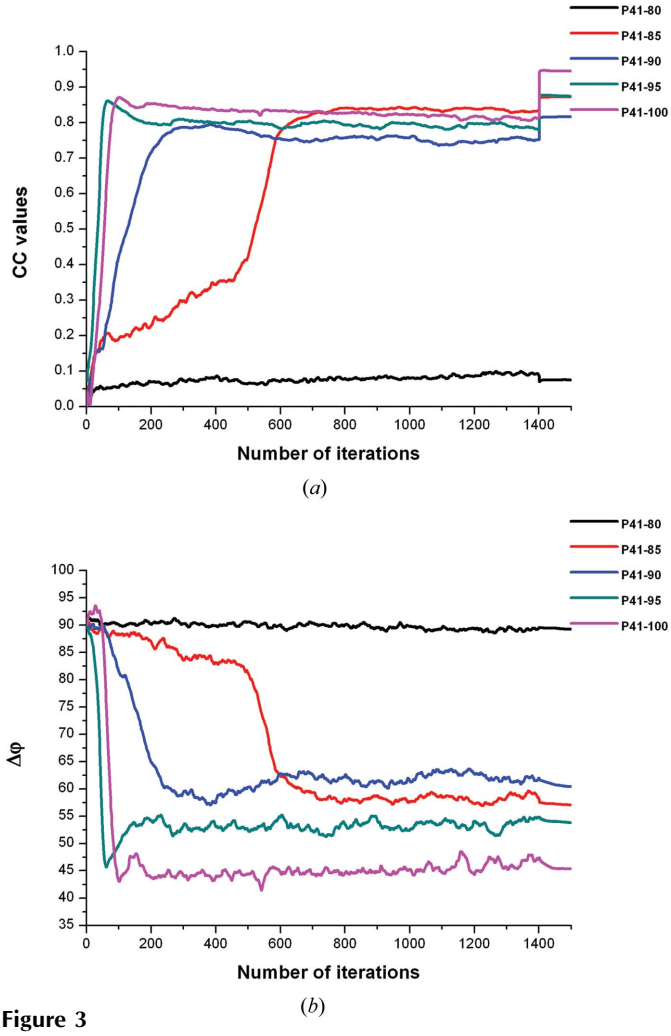


Figure 3
The convergent processes of P41-100, 95, 90, 85, 80 after 1400 iterations of the phasing algorithm and the final stage of 100 cycles of solvent flattening with histogram matching. (*a*) CC values plotted *versus* the number of iterations, (*b*) average errors in phase angles $\Delta\varphi$ plotted *versus* the number of iterations. P41-85, 90, 95, 100 in red, blue, dark cyan, magenta, respectively, all have convergent results. P41-80 in red has no obvious convergent process.

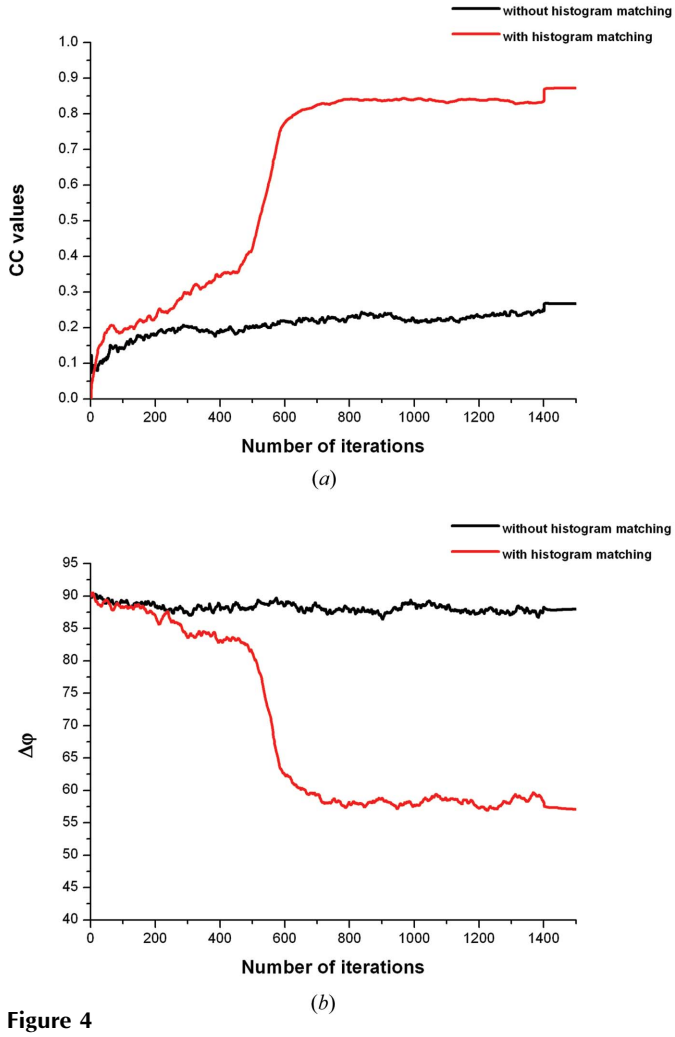


Figure 4
The convergent process of P41-85 which has a solvent content of 52.25% after 1400 iterations of the phasing algorithm and the final stage of 100 cycles of solvent flattening without histogram matching (in black) and with histogram matching (in red). (*a*) CC values plotted *versus* the number of iterations, (*b*) average errors in phase angles $\Delta\varphi$ plotted *versus* the number of iterations.

it (in black). The final CC value is 0.87 when applying histogram matching and 0.27 when not applying it, $\Delta\varphi$ converges to 57.0° after applying histogram matching and to 87.9° when not applying it. After applying histogram matching the density map in the case of P41-85 is clearer and easier to interpret than before, as illustrated in Fig. 5. We believe that histogram matching, as a real-space constraint in the iterations, can lower the demand in solvent contents and enhance the efficiency of this method. The results of the different cases are illustrated in Table 1.

3.3. Insensibility to errors

There are errors associated with the measurement of diffraction data $I(h, k, l)$ collected by X-ray detectors. To examine the applicability of this approach to experimental data, we have to study the sensitivity of the algorithm to errors. We simulated the effects of errors in our computer modelling. Making the assumption that all the errors reside in $|F(h, k, l)|$ and that errors follow a Gaussian distribution, the probability of $I_p(h, k, l)$ having a certain value is then

$$I_p(h, k, l) \propto \exp[-(\varepsilon - am)^2/2\sigma^2],$$

where am is the mean value, σ the standard deviation. In this instance, am is equal to $I(h, k, l)$ and σ is selected as $10 \times |F(h, k, l)|$ and $100 \times |F(h, k, l)|$ which means 10 times and 100 times $|F(h, k, l)|$, termed $10F$ and $100F$, respectively.

As is shown in Fig. 6, we added errors of $10F$ in the diffraction intensities in the case of P41-85 (in black); after 1400 iterations of the phasing algorithm and the final stage of 100 cycles of solvent flattening our algorithm successfully obtains a convergence result with the final CC = 0.86 and $\Delta\varphi$ = 57.1°. We then add errors of $100F$ in the diffraction intensities (in red). A good reconstruction result with the final CC = 0.84 and $\Delta\varphi$ = 59.2° is also shown in Fig. 6. The convergent density map of applying $100F$ of errors in Fig. 7 clearly displays the structures of the proteins. The reason for the insensibility to errors is that the electron densities are the result of the combination of all the diffraction data observed in experiments; although the Gaussian distribution error might be very

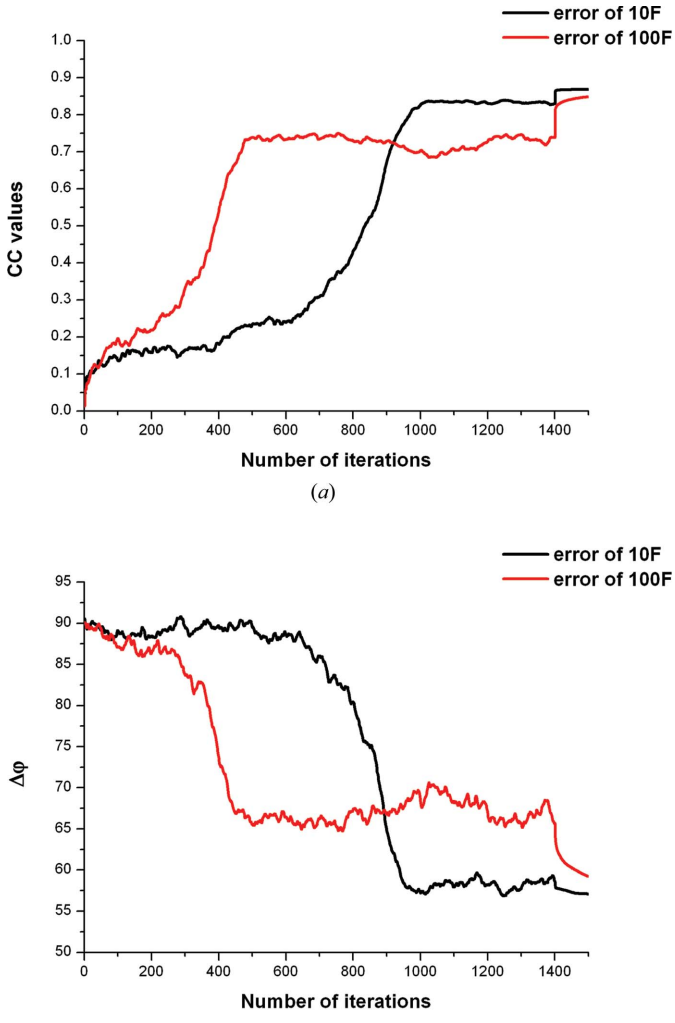


Figure 6
The convergent process of P41-85 which has a solvent content of 52.25% after 1400 iterations of the phasing algorithm and the final stage of 100 cycles of solvent flattening with an error of $10F$ (in black) or $100F$ (in red). (*a*) CC values plotted *versus* the number of iterations, (*b*) average errors in phase angles $\Delta\varphi$ plotted *versus* the number of iterations.

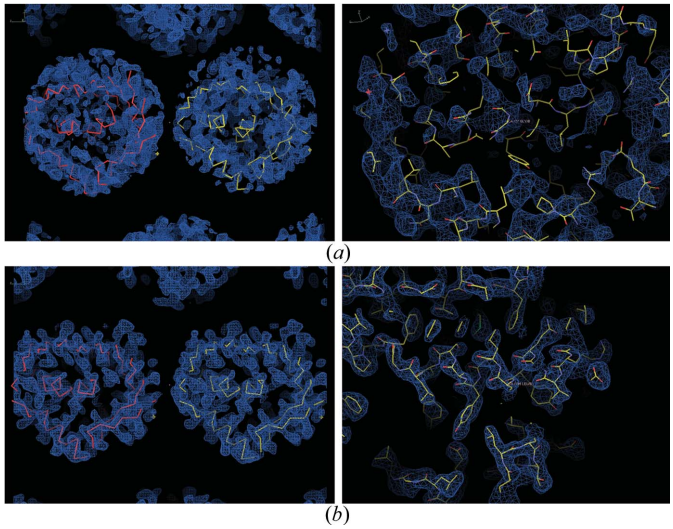


Figure 5
(*a*) The 2.0 Å electron-density map of P41-85 after convergence without histogram matching, and (*b*) with histogram matching, compared with the α-carbon tracing of the constructed model (left) and the final structures superimposed (right).

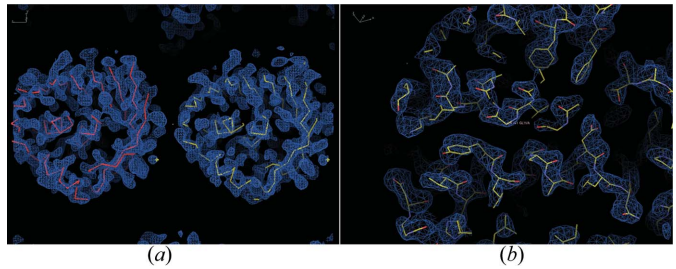


Figure 7
(*a*) The 2.0 Å electron-density map of P41-85 after convergence (in blue) with added error of $100F$, compared with the α-carbon tracing of the constructed model superimposed. (*b*) A small section of the map with the final structure superimposed.

big in each diffraction peak, this kind of spontaneous error should be alleviated during the Fourier synthesis. Since our iterative process is based on the Fourier transformation rather than the amplitude relationship in the direct method, the precision of the amplitude is not that crucial any more.

3.4. The influence of resolution

A high resolution of a three-dimensional structure is necessary to understand the functions of proteins at a molecular level. We therefore tested the influence of different resolutions on the phasing algorithm. There are five cases derived from P41-85 with resolutions ranging from 1.6 to 3.2 Å in 0.4 Å steps. Shown in Fig. 8, after 1400 iterations of the phasing algorithm and the final stage of 100 cycles of solvent flattening, in the cases of 1.6, 2.0 and 2.4 Å there is a convergence result with a high CC above 0.8 and $\Delta\varphi$ less than 65°. However there are some instable fluctuations in the cases of 2.8 and 3.2 Å, and the final result is not satisfactory as shown in Fig. 8 in dark cyan and magenta, respectively. A possible reason is that the constraints in real space are gradually weakened as the resolution decreases. When the resolution of a three-dimensional structure is around 3.0 Å, there are some ambiguities in the protein secondary structure; this may cause an accumulative discrepancy in each iteration which means the phasing algorithm does not function very well.

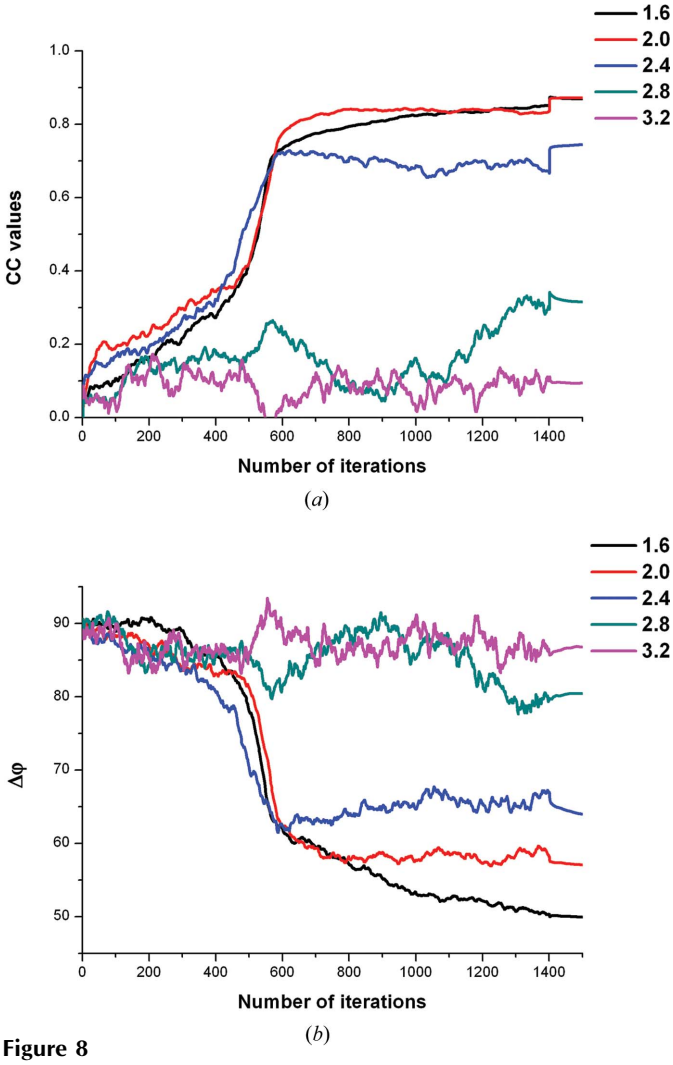


Figure 8
The convergent process of P41-85 with resolutions ranging from 1.6 to 3.2 Å in 0.4 Å steps. (*a*) CC values plotted *versus* the number of iterations, (*b*) average errors in phase angles $\Delta\varphi$ plotted *versus* the number of iterations. The cases of 1.6, 2.0 and 2.4 Å in black, red, blue, respectively, have convergent results. The cases of 2.8 and 3.2 Å in dark cyan and magenta, respectively, have unstable fluctuations in the evolution of reconstruction.

3.5. Reconstructing near-forward low-resolution data

In reality, it is impossible to obtain the full experimental data including all of the near-forward low-resolution data because of the beam stop. Unfortunately this part of the data is important. Many people have pointed out that the HIO phase-retrieval algorithm is quite sensitive to near-forward low-resolution data (or missing central data) (Fienup & Wackerman, 1986). The near-forward low-resolution data confer the shapes of proteins, namely the optimized density regions D. If the missing near-forward low-resolution data are serious, there is no rigorous constraint upon the correct shapes of proteins in the diffraction data, and the iterative process would fall into a false minimum. Thus, we reconstruct the missing near-forward low-resolution data as follows:

Regions E represent the data that we can detect, otherwise E' stands for the missing regions including the near-forward low-resolution data. During the iterations, in the mth cycle, after modifying $\rho'_m(x, y, z)$ according to equation (2), we obtain a full set of complete data $|F_m^{\rm calc}(h, k, l)|$ from Fourier transformation; then we should replace $|F_m^{\rm calc}(h, k, l)|$ by $|F^{\rm exp}(h, k, l)|$ in the region E. Unfortunately, the values of $|F^{\rm exp}(h, k, l)|$ in the region E' are not available; therefore we use $|F_m^{\rm calc}(h, k, l)|$ in the region E', or in other words $|F_m^{\rm calc}(h, k, l \in E'))|$, to calculate the lost $|F^{\rm exp}(h, k, l)|$, by multiplying by a scale factor which is the quotient of dividing the sum of $|F^{\rm exp}(h, k, l)|$ by the sum of $|F_m^{\rm calc}(h, k, l)|$ in the regions E. That is to say, after m iterations, we calculate $|F_m^{\rm exp}(h, k, l \in E')|$ by the following equation (6):

$$\left|F_{m\ hkl\in E'}^{\ \ \rm exp}(h, k, l)\right| = \left|F_{m\ hkl\in E'}^{\ \ \rm calc}(h, k, l)\right| \times \frac{\sum_{hkl\in E}\left|F^{\rm exp}(h, k, l)\right|}{\sum_{hkl\in E}\left|F_{m}^{\ \rm calc}(h, k, l)\right|}. \quad (6)$$

These reconstructed data are not of constant value and may change in each iteration. By employing this method, we can get a full set of $|F^{\rm exp}(h, k, l)|$ and can use the algorithm for successive iterations. In later discussion, we use an example to show how this is executed (see Fig. 11).

4. Phasing from the envelopes of some high-solvent-content protein crystals

4.1. Molecular shape determination in terms of spherical harmonics

SAXS is a method developed recently for determining the shapes of biological macromolecules in solution. From the

Table 2
The f_{lm} coefficient distribution at the precision of l = 5 calculated by *CRYSOL* for the case of 1y5y, where Z_0 = 31.146578.

Envelope function from 1y5y.

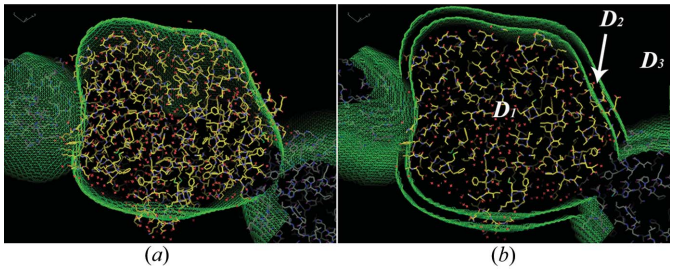
l	m	Real part of f_{lm}	Imaginary part of f_{lm}
0	0	3.417654	0.000000
1	0	−0.008363	0.000000
1	1	−0.041268	−0.013553
2	0	0.242316	0.000000
2	1	−0.143523	−0.167466
2	2	0.016771	−0.322748
3	0	0.021200	0.000000
3	1	−0.010219	0.073997
3	2	−0.016515	−0.015955
3	3	0.005789	−0.046974
4	0	0.139274	0.000000
4	1	0.054295	0.031538
4	2	0.012270	0.000759
4	3	0.060709	−0.069023
4	4	0.099641	0.004127
5	0	0.058214	0.000000
5	1	−0.048134	0.028456
5	2	0.021700	0.013332
5	3	−0.003628	−0.043533
5	4	−0.023510	0.022587
5	5	0.023720	0.015379

SAXS data, one can determine the biological molecular envelope by a two-dimensional angular function $\omega(\theta, \varphi)$ describing the molecular boundary such that the particle density $\rho(r)$ is unity inside and vanishes elsewhere. The function $\omega(\theta, \varphi)$ can conveniently be expanded into a series of spherical harmonics $Y_{lm}(\theta, \varphi)$ according to the following equation:

$$\omega(\theta, \varphi) = Z_0 \sum_{l=0}^{L} \sum_{m=-l}^{l} f_{lm} Y_{lm}(\theta, \varphi), \qquad (7)$$

with f_{lm} being complex multipole coefficients and l representing the multipole order. Z_0 is a scale factor. Furthermore, where $P_l^m(\cos\theta)$ are associated Legendre functions (with argument $\cos\varphi$), and l and m are integers with $-l \leq m \leq l$ (Svergun & Stuhrmann, 1991; Svergun *et al.*, 1996)

$$Y_{lm}(\theta, \varphi) = \left[\frac{(2l+1)(l-m)!}{4\pi(l+m)!}\right]^{1/2} P_l^m(\cos\theta)\exp(im\varphi). \qquad (8)$$



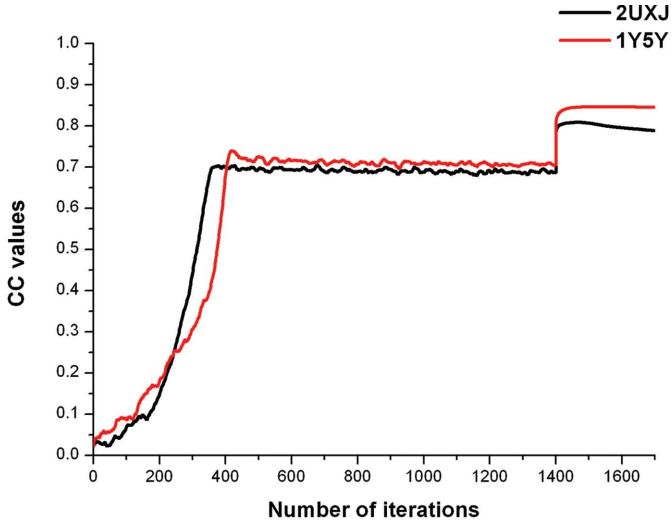
(*a*) (*b*)

Figure 9
(*a*) The envelope of 1y5y after expanding according to the symmetries of space group $P4_32_12$. The envelope is shown in green. (*b*) The envelope of 1y5y which has a block region D_2, as shown in green.

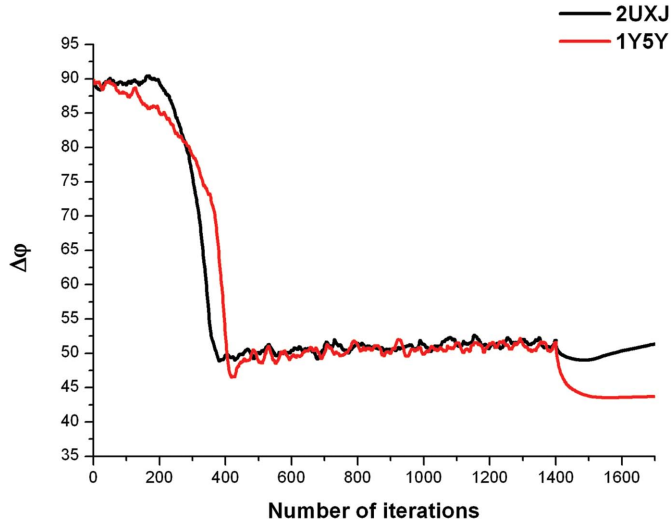
Table 3
The five cases downloaded from the PDB and the final results of sufficient iterations.

ID	Space group	Solvent content (%)	Resolution (Å)	Final CC	$\Delta\varphi$ (°)
1y5y	$P4_32_12$	68.00	2.00	0.85	45.2
2uxj	$P4_32_12$	76.56	2.25	0.79	50.9
3iai	$P6_1$	77.89	2.20	0.72	54.0
1n0h	$P422$	65.72	2.80	0.69	58.2
2hnk	$P222_1$	69.86	2.30	0.60	69.3

Svergun *et al.* (1995) developed a program named *CRYSOL* for evaluating the solution scattering from macromolecules with known atomic structures; thus, we downloaded some high-solvent-content protein structures from the PDB and used *CRYSOL* to predict the SAXS data (Svergun *et al.*,



(*a*)



(*b*)

Figure 10
The convergent process of 2uxj and 1y5y after 1400 iterations of the phasing algorithm and the final stage of 300 cycles of solvent flattening. (*a*) CC values plotted *versus* the number of iterations, (*b*) average errors in phase angles $\Delta\varphi$ plotted *versus* the number of iterations.

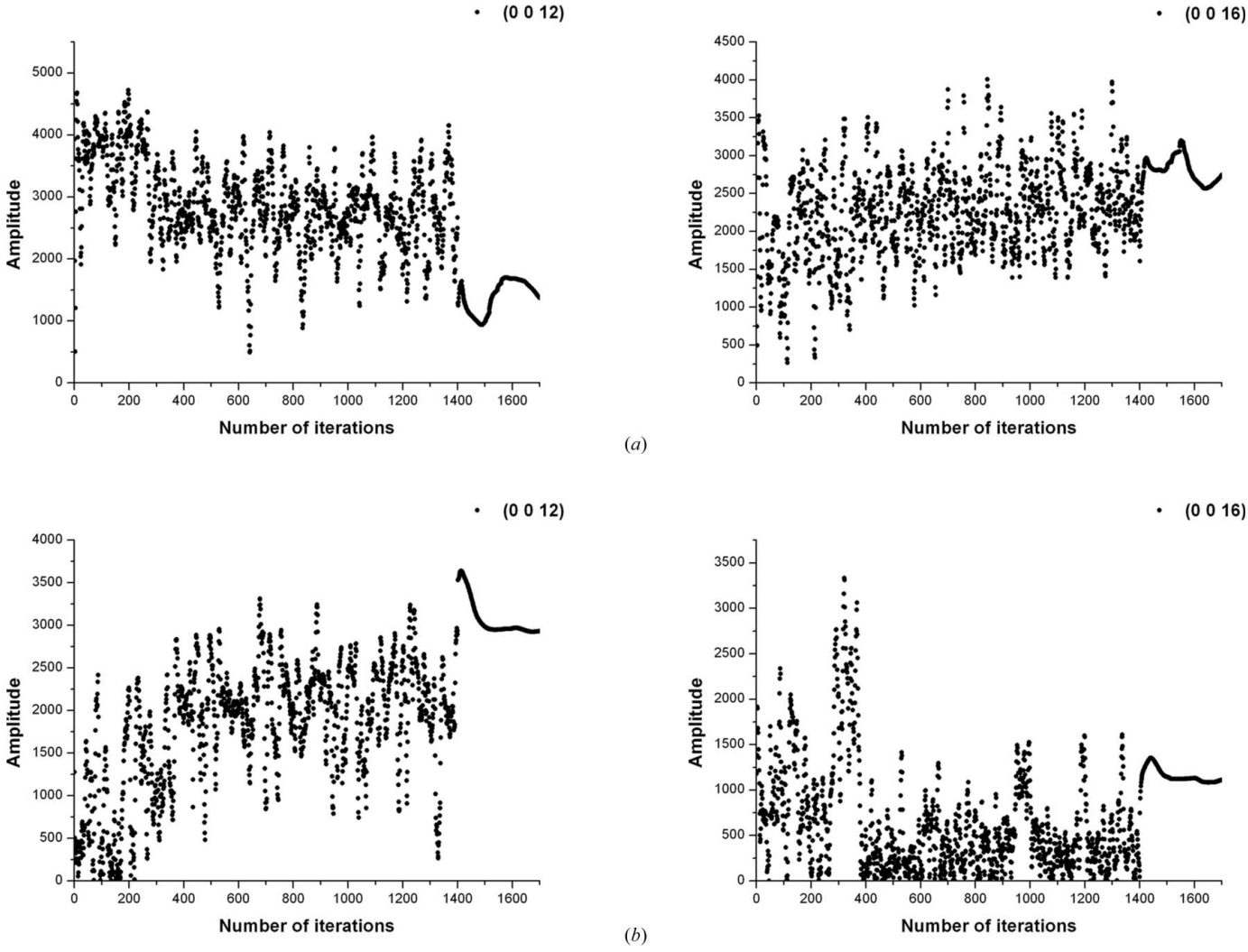


Figure 11
The missing amplitudes of reflections 0,0,12 and 0,0,16 in (*a*) 2uxj and (*b*) 1y5y reconstructed from equation (6) *versus* the number of iterations.

1995). The factors f_{lm} are complexes, the real part and imaginary part of f_{lm} are shown in Table 2. According to equations (7), (8) and (9),

$$\rho(r) = \begin{cases} 1, 0 \le r \le \omega(\theta, \varphi) \\ 0, r > \omega(\theta, \varphi) \end{cases}, \tag{9}$$

we can get the envelopes of the protein molecules. Using the envelopes and expanding by the symmetries of space groups, we shall know the non-density regions occupied by solvents and the density regions where the proteins are located in the cells. An example of the envelope of protein 1y5y from the PDB is illustrated in Fig. 9(*a*).

4.2. Adding a block region

In Fig. 9(*a*), the protein molecules are basically wrapped by the envelopes, but there are still a small number of residues of the polypeptide chains out of the envelopes. In our tests, without appropriate treatment for these special residues, the iterative process would fall into a false minimum. To solve this stagnation problem, we introduced an improvement in the third step of the method, which ignores the effect of these special residues; the details are given below:

Separate the unit cell into three parts: (i) the density regions D_1 where the proteins are located; (ii) the block shell D_2 surrounding the molecules, usually 3 or 4 Å thickness; (iii) the remaining regions D_3 occupied by the solvents. An example of 1y5y with a block region is shown in Fig. 9(*b*). Then in the third step of the method, we modified $\rho_m(x, y, z)$ according to equation (10) instead of equation (2),

$$\rho_m(x, y, z) = \begin{cases} \rho_{m-1}(x, y, z), [(x, y, z) \in D_1] \\ 0, [(x, y, z) \in D_2] \\ \rho_{m-1}(x, y, z) - \varepsilon\rho'_m(x, y, z), [(x, y, z) \in D_3] \end{cases}, \tag{10}$$

which pushed the grid points inside the block region D_2 always to zero in the iterative process, and pushed the grid points in the region D_3 close to zero to fulfil the real-space constraint based on the HIO. Since the block region may have a few electron densities, we force this region to be always zero to weaken the effect of this region for the whole iterative process.

4.3. Results and discussion

After the improvement (applying histogram matching, reconstructing near-forward low-resolution data and adding a block region) in this method, we have performed phase retrievals for several cases of protein crystals with the amplitudes of structure factors derived from diffraction data collected by X-ray detectors and envelopes calculated from the structure files downloaded from PDB files. Table 3 illustrates the final results of five cases in descending order of CC values. The space groups are different in the five cases and resolutions range from 2.0 to 2.8 Å; in each case the crystal cell has a high solvent content above 65% to satisfy the oversampling condition. The final CC values in the five cases are larger than 0.60 and the values of $\Delta\varphi$ are smaller than 70°. Such values mean that the density maps are able to trace the polypeptide chains of the proteins.

Shown in Fig. 10, 1y5y (in red) and 2uxj (in black) have solvent contents of 68.00 and 76.56%, respectively, and after the same 1400 iterations of the phasing algorithm and the final stage of 300 cycles of solvent flattening, they both converge with CC = 0.85 and 0.79, $\Delta\varphi$ = 45.2 and 50.9°, respectively. For the two cases, we monitored the reconstruction of missing low-resolution data. As shown in Fig. 11, the amplitudes of reflections 0,0,12 and 0,0,16 reconstructed from equation (6) fluctuated wildly and irregularly in the first part of the 1400 iterations and vibrated only slightly in the last 300 iterations. Since the iterative process is based on the Fourier transformation rather than the amplitude relationship in the direct method, the precision of the amplitude is not that crucial, but the existence of values for these reflections does help the iterations to avoid falling into a false minimum. When convergences were reached after 1400 cycles, we modified the solvent region to zero, so the amplitudes of missing low-resolution reflections were more stable than before. This caused a little vibration of the CC and $\Delta\varphi$ curves in the convergence result, as shown in the case of 2uxj. In our test, once a convergence result is obtained after the first part of sufficient iterations, applying solvent flattening a few times rather than hundreds to thousands of times is adequate.

The reconstructed density map is shown in Fig. 12. In cases such as 1y5y and 2uxj, the final CC reaches 0.85 and 0.79, and the average error in the phases $\Delta\varphi$ is as small as 45.2 and 50.9°, respectively. One can clearly see the connectivity of the main chain and the fit of the side chain at the resolutions of the data sets (2.00 Å for 1y5y, 2.25 Å for 2uxj) from Figs. 12(*a*) and 12(*b*), which provide a high-quality density map for structure determination. Because of the block region, small parts of densities at the edge of the envelope are missing, but this does not have a great influence on the map interpretation. For the case of 2hnk, the final CC = 0.60 and $\Delta\varphi$ = 69.3°, and the quality of the density map shown in Fig. 12(*c*) at 2.30 Å resolution is poor in comparison with the other two cases discussed below; however, one can still see the connectivity of the main chain and the fit of the side chain from the density map. These maps are a good starting point for *ab initio* structure determination. It should be noted that the convergence speed is somewhat different in each case. 3iai reaches the convergence result after 400 iterations and a final 100 cycles of solvent flattening, whereas 2hnk and 1n0h need more iterations (2900 iterations and 100 cycles of solvent flattening, and 7500 iterations and 100 cycles of solvent flattening, respectively). Because of the different qualities of the initial X-ray diffraction data, it is natural that the convergent processes are different in each case.

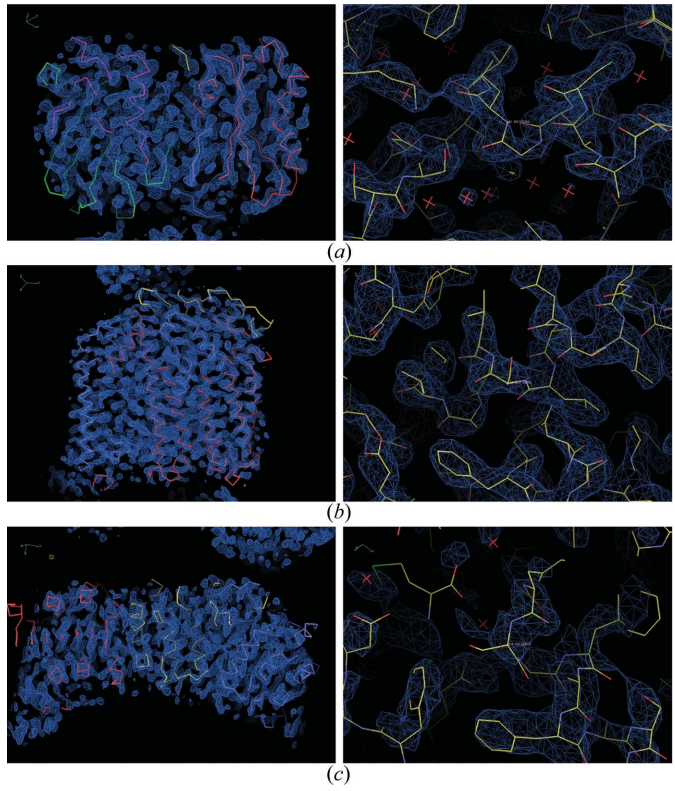


Figure 12
The reconstructed electron-density map after sufficient convergence. (*a*) 2.00 Å map of 1y5y, (*b*) 2.25 Å map of 2uxj, (*c*) 2.30 Å map of 2hnk, compared with the α-carbon tracing of the constructed model (left) and the final structure superimposed (right).

5. Conclusions

It has been demonstrated that the phases of reflections can be retrieved using this iteration method, which appears to have a fair degree of robustness against the errors in the intensity data; however, it requires a high solvent content in the protein crystals. The cases illustrated in Table 3 all have a high solvent content, more than 65%; hence the oversampling condition is satisfied and the phases can be solved by this method. Reflection data with a resolution better than 3.0 Å are favorable for this algorithm since they meet the requirements for distinguishing the secondary structures of proteins. It is anticipated that other density-modification methods, such as non-crystallographic symmetry averaging, could break through this restriction. It is hoped that this method will greatly facilitate the *ab initio* structure determination of proteins.

This work was supported by grants from the National Natural Science Foundation of China (grant No. 10979005) and the National Basic Research Program of China (grant No. 2009CB918600).

References

Abrahams, J. P. (1997). *Acta Cryst.* D**53**, 371–376.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Cowtan, K. D. & Main, P. (1996). *Acta Cryst.* D**52**, 43–48.
Ealick, S. E. (1997). *Structure*, **5**, 469–472.
Fienup, J. R. (1982). *Appl. Opt.* **21**, 2758–2769.
Fienup, J. R. & Wackerman, C. C. (1986). *J. Opt. Soc. Am. A*, **3**, 1897–1907.
Hao, Q. (2001). *Acta Cryst.* D**57**, 1410–1414.
Hendrickson, W. A., Smith, J. L., Phizackerley, R. P. & Merritt, E. A. (1988). *Proteins*, **4**, 77–88.
Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H. & Phillips, D. C. (1958). *Nature* (*London*), **181**, 662–666.
Lunin, V. Yu. (1988). *Acta Cryst.* A**44**, 144–150.
Lunin, V. Y., Lunina, N. L., Petrova, T. E., Skovoroda, T. P., Urzhumtsev, A. G. & Podjarny, A. D. (2000). *Acta Cryst.* D**56**, 1223–1232.
Lunin, V. Yu. & Skovoroda, T. P. (1991). *Acta Cryst.* A**47**, 45–52.
Lunin, V. Yu., Urzhumtsev, A. G. & Skovoroda, T. P. (1990). *Acta Cryst.* A**46**, 540–544.
Lunin, V. Yu. & Vernoslova, E. A. (1991). *Acta Cryst.* A**47**, 238–243.
Miao, J., Kirz, J. & Sayre, D. (2000). *Acta Cryst.* D**56**, 1312–1315.
Miao, J. & Sayre, D. (2000). *Acta Cryst.* A**56**, 596–605.
Murthy, H. M., Hendrickson, W. A., Orme-Johnson, W. H., Merritt, E. A. & Phizackerley, R. P. (1988). *J. Biol. Chem.* **263**, 430–436.
Perutz, M. F. (1956). *Acta Cryst.* **9**, 867–873.
Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.
Svergun, D., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.
Svergun, D. I. & Stuhrmann, H. B. (1991). *Acta Cryst.* A**47**, 736–744.
Svergun, D. I., Volkov, V. V., Kozin, M. B. & Stuhrmann, H. B. (1996). *Acta Cryst.* A**52**, 419–426.
Wang, B. C. (1985). *Methods Enzymol.* **115**, 90–112.
Weeks, C. M. & Miller, R. (1999). *J. Appl. Cryst.* **32**, 120–124.
Zhang, K. Y. J. & Main, P. (1990*a*). *Acta Cryst.* A**46**, 41–46.
Zhang, K. Y. J. & Main, P. (1990*b*). *Acta Cryst.* A**46**, 377–381.