Real-Space Envelopes by the Connectivity Constraint: An Approach to the Ab initio Macromolecular Phase Problem

Guangzhou Zou, George N. Phillips, Jr.

CRPC-TR97743
August 1997

Center for Research on Parallel Computation
Rice University
6100 South Main Street
CRPC - MS 41
Houston, TX 77005

Submitted May 1998

Real-Space Envelopes by the Connectivity Constraint: An Approach to the Ab initio Macromolecular Phase Problem

Guangzhou Zou George N. Phillips, Jr.

August 15, 1997

Abstract

An approach is proposed for the ab initio macromolecular phase problem in which the information that proteins are make up of connected chains of atoms is exploited by connectivity constraints. It is show here that the simple forms of the connectivity constraint together with the non-negativity constraint can generate approximate molecular envelopes from the diffraction data $|Fhkl|$. Examples with different solvent contents, space groups, and numbers of molecules per asymmetric unit cell were chosen to illustrate the applicability of the algorithms.

1 Introduction

1.1 The Phase Problem

The phase problem (Hauptman, 1989) arises because a monochromatic diffraction experiment on a single crystal provides only the amplitudes of the reflections. The phases of the reflections cannot be recorded in the experiment. However, both the phase and the amplitude of each reflection are required to recover the electron density in the crystal by the inverse Fouier transformation. Thus, the phase problem can be viewed as the recovery of an object from the amplitudes of its Fourier transform.

The phase problem has been solved in the case of small molecules (up to a few hundred atoms in the unit cell) through the so-called direct methods (Karle & Hauptman, 1956). However, the application of these direct methods to macromolecules has not yet succeeded.

1.2 The uniqueness of the macromolecular phase problem

The question of the uniqueness of the solution in the phase problem is important because even the most sophisticated searching algorithms is bound to fail when the true solution is only one of a very large number of distributions satisfying all of the constraints.

1.2.1 Phase recovery in in optics

In optics, an unique solution exists for the phase problem in almost all cases if the object is sampled continuously in two or higher dimensions (Bates, 1982; Hayes, 1982). In these cases, the uniqueness depends on whether the structure factor

$$F = \sum_{x,y} \rho(x,y) exp[2\pi i(hx + ky)] \tag{1}$$

is factorizable. In one-dimensional problems, F is polynomial in a single variable, and the fundamental theorem of algebra guarantees that F can always be expressed as a product of first-order polynomials. Thus, the phase problem in one dimension generally does not have a unique solution. However, there is no equivalent of the fundamental theorem in higher dimensions, and in fact almost all polynomials in two or higher dimensions are irreducible (Hayes, 1987). This has been borne out in practice by the development of simple iterative algorithms for reconstructing an object from only the magnitude data (Dainty & Fienup, 1987).

1.2.2 Phase problem in x-ray crystallography

The application of the iterative algorithms to crystallography fails because the Fourier transform of the object can only be sampled at discrete points in the crystallographical experiments. For the continuous function

$$I = FF*, \tag{2}$$

where * indicates the complex conjugate, to be completely specified by its values at discrete points, it must at least be sampled at the Shannon limit. For simplicity, let's consider an one-dimensional example. The intensity of reflection I is seen to be the transform of the autocorrelation of object function $\rho(x)$:

$$I = \sum_{x=-L}^{L} \left[\sum_{x'=-L/2}^{L/2} \rho(x')\rho(x+x') \right] exp(2\pi ihx). \tag{3}$$

the limits on the outer sum are twice that of the inner sum since with ρ non-zero for $-L/2 < x' < L/2$, the autocorrelation of ρ will be non-zero for $-L < x < L$. As the spectral width of I is twice that of F, the Nyquist spacing required for I is one half that required for F. But, the periodicity of the crystal restricts sampling to integer multiples of 1/L, which suffices for F but not for I. The polynomial I is thus not uniquely specified by crystal diffraction data, and hence there is no equivalent of

$$I = FF*$$

to uniquely determine F in crystallography. Additional chemical information is clearly required to uniquely define ρ, but the amount needed is unknown (Milane, 1990; Baker *et al.* 1993).

1.2.3 Connectivity constraint

Simple constraints from chemical considerations such as atomicity and positivity are sufficient for small molecule crystal structures at atomic resolution to solve the phase problem. In small-molecule crystallography, the number of measured $|Fhkl| \gg 3(N-1)$, where N is the number of atoms. Thus, the constraints of atomicity and positivity makes the structure-determination problem greatly overdetermined. For proteins, however, this is generally not the case. Stronger chemical constraints will be required to compensate for the low ratio of data to free parameters intrinsic to the macromolecular phase problem.

Notice that the macromolecular phase problem become greatly overdetermined once it is possible to trace an atomic model through the density and thus make use of the detailed rules of connectivity in stereochemistry. Thus, the detailed connectivity represents a strong chemical constraint that contains enough information to specify an unique solution for the macromolecular phase problem. Unfortunately, in general this constraint can only by utilized after an approximate model is determined.

However, even without an approximate model, the information that proteins are made up of connected chains of atoms can be exploited at relatively early stages in the solution of macromolecular structure. Wilson and Agard (1993), Baker *et al* (1993) and Bystroff *et al.* (1993) approximated the connectivity of the protein by skeletonization of the electron density map. The skeleton was thinned and pruned by modification of the algorithm described by Greer (1985). This approximate connectivity constraint can become very powerful and start to provide significant amount of information in the situation where the the phase error is less than 50°.

Since the connectivity is such a powerful constraint in general, it would be worthwhile to find out if there is some sort of connectivity constraint that can be employed at the earliest stages in the ab initio phasing and how much information can be provided by the connectivity constraint. In this paper, we will show that a coarse connectivity constraint can be applied at the very beginning of a searching process in which the solution is searched by optimizing its agreement with observed diffraction intensities. The additional information provided by the constraint specifies real-space envelopes for the macromolecules.

2 Definitions

In this approach, the electron-density map is calculated on 3-dimensional regular grid determined by the resolution of the measured data.

- The *neighborhood* of a grid point (x, y, z) is defined as a set of 27 grid points: $(x + d, y + d, z + d)$ with $d = -1, 0,$ or 1.
- The nearest neighbors of a grid point (x, y, z) is defined as $(x \pm 1, y, z)$, $(x, y \pm 1, z)$, $(x, y, z \pm 1)$;
- A grid point is *marked* if it is considered as a part of the molecule and is *unmarked* otherwise.
- The marked points are considered to *connected* if there is a continuous path of marked nearest-neighbors between them.

- A *cluster* of marked points is a set of marked points that are mutually connected.

- The *size of a cluster* is defined as the number of marked points in the cluster.

- An *isolated point* is a marked grid point and also the only marked point in its neighborhood.

3 Iterative Algorithm

In this section, we describe the basic iterative algorithm under a simple connectivity constraint and the application of this algorithm to the measured diffraction data of tropomyosin molecule.

3.1 The iteration protocol

The iterative transform algorithm is also known as the error-reduction algorithm has been widely used in the optical image recovery (Stark, 1987). The error-reduction algorithm was chosen because it enable the easy input of connectivity information from the real space. The algorithm we proposed here consists of the following five steps (for nth iteration):

1. Fast Fourier Transform (FFT) of $f_n(x, y, z)$, an estimated density map $\rho(x, y, z)$, yielding $G_n(h, k, l)$;

2. scale the $G_n(h, k, l)$ with respect to the measured $|F|$;

3. replace the magnitude of $G_n(h, k, l)$ with the measured data if the data is available which allow it to satisfy the Fourier-domain constraints to form $G'_n(h, k, l)$, an estimate of $F(h, k, l)$;

4. inverse FFT of $G'_n(h, k, l)$, yielding $f'_n(x, y, z)$;

5. apply a real-space filter to $f'_n(x, y, z)$ to form $f_{n+1}(x, y, z)$.

The real-space filter modifies $f'_n(x, y, z)$ to allow it to satisfy the constraints of positivity, connectivity and symmetry.

The function $f'_n(x, y, z)$ is defined in a grid space. We assume that the molecule can be approximated by N grid points with non-zero density values in the asymmetic unit. The value N, which is much less than the total number of grid point, is a user defined parameter.

3.2 Real-space filter-0

A real-space filter imposing only positivity constraint consists of two steps:

1. set the density value of unmarked point be zero; and
2. if the density value of a marked point is less than zero, set it to zero.

We call this filter the real-space filter-0.

3.3 Real-space filter-1

A simple real-space filter with the consideration of connectivity, which we call real-space filter-1, can be implemented as following:

1. sort the grid points according to their density values in the neighborhoods of the marked points in the asymmetric unit;
2. mark the first N grid points with the highest density values and unmark the rest grid points in the asymmetric unit;
3. unmark any isolated points;
4. set the density value of unmarked point be zero;
5. if the density value of a marked point is less than zero, set it to zero; and
6. use the symmetric operations to extend above modifications to the whole unit cell.

3.4 Random method for initial density generating

The initial estimate of the density map $f_0(x, y, z)$ can be generated by a random method:

1. randomly place N marked points in the asymmetric unit;
2. set the density values of marked points to 1.0 and unmarked points to zero; and
3. apply symmetric operations to create a map $f_0(x, y, z)$ over the entire unit cell.

3.5 Application example

Tropomyosin molecules are composed of alpha helices that are about 284 amino acids in length. The structure was solved initially at 9 Åresolution. The tropomyosin molecules crystallize in space group C2. The crystals have unit cell parameters $a = 259.7$, $b = 55.3$, $c = 136.3$, $\alpha = \gamma = 90.0$, and $\beta = 97.2$. The resolution of the diffraction data ranges from 2.233 Åto 87.7 Å. The unit cell contains $128 \times 32 \times 64$ grid points. We used $N = 10,000$ grid points to approximate the electron density in the asymmetric unit. Only 95% of the reflection were used in the iteration precess and the other 5% were used for calculating free r values. The results are show in Fig. 1.

It seems that the algorithm can quickly digest the connectivity information contained in the real-space filter-1 and produce approximate envelopes for certain types of molecules. However, the error-reducing process becomes saturated rapidly and the free r cannot be improved by the large number of iterations. At this stage, it seems that additional information of connectivity is needed in order to reduce the phase error further.

4 Alternative algorithms

In this section, we discuss two alternative algorithms in which higher connectivity constraints are imposed. The basic iteration protocol will remain the same. However, different real-space filters will be constructed. Also, the initial density map will be generated differently.

4.1 Clustering method for initial density generating

The random method discussed previously generates the initial map by a pure random process. The new method specified here will create the starting density map with the consideration of connectivity of the map. The method consists of the following procedures:

1. generate a random initial density map $f_0(x, y, x)$ by the random method described previously;
2. perform one iteration using real-space filter-1 to get a new density map $f_1(x, y, z)$;
3. find the clusters formed by the marked grid points in the asymmetric unit and sort the clusters with respect to their sizes;
4. keep the first three largest clusters and unmark the grid points that belongs to other clusters;
5. set the density value of the unmarked points to 0.0;
6. create a new density map $f_1'(x, y, z)$ over the entire unit cell by symmetric operations;
7. count the total number of the marked points, P, corresponding to the density map $f_1'(x, y, z)$;
8. repeat steps (1) to (7) described above many times, say 1,000 times, and select the $f_1'(x, y, z)$ that has the maximum number of marked points, P;
9. FFT of $f_1'(x, y, z)$ to get a function $G_1'(h, k, l)$ in reciprocal space.
10. scale the $G_1'(h, k, l)$ with respect to the measured $|F|$;
11. replace the magnitude of $G_1'(h, k, l)$ with the measured data if the data is available to form $G_1''(h, k, l)$;
12. inverse FFT $G_1''(h, k, l)$, yielding $f_1''(x, y, z)$;
13. Sort the unmarked points in the asymmetric unit according to their density values.

14. Add $D = 1.5N - P$ marked points to the asymmetric unit such that there are total of 1.5N marked point in the asymmetric unit. The marked points are added in the following way: mark $0.8D$ unmarked points that have lowest density values and mark $0.2D$ unmarked points that have highest density values.

15. Set the density value of unmarked point be zero;

16. If the density value of a marked point is less than zero, set it to zero;

17. Use the symmetric operations to extend above modifications to the whole unit cell to form the initial map.

4.2 Real-space filter-2

The real-space filter-2 will put a stronger connectivity constraint on the estimated density map than real-space filter-1. The new filter will have better performance under certain circumstance. The real-space filter-2 is implemented as follows.

1. Sort the grid points according to their density values in the neighborhoods of the marked points in the asymmetric unit.

2. Mark the first N grid points with the highest density values and unmark the rest grid points in the asymmetric unit.

3. Find the clusters formed by the newly marked grid points and sort the clusters according to their sizes.

4. Keep the grid points in the first M largest clusters marked and unmark the rest grid points.

5. Count the number of the marked points, P, in the asymmetric unit.

6. Sort the unmarked points in the asymmetric unit according to their density values.

7. Add $D = 1.5N - P$ marked points to the asymmetric unit such that there are total of 1.5N marked point in the asymmetric unit. The marked points are added in the following way: mark $0.8D$ unmarked

points that have lowest density values and mark $0.2D$ unmarked points that have highest density values.

8. Set the density value of unmarked point be zero;
9. If the density value of a marked point is less than zero, set it to zero;
10. Use the symmetric operations to extend above modifications to the whole unit cell.

Most of the newly marked points in step (7) are low density points because this would keep the change of the map to the minimum. Once a point becomes marked, its density value is no longer restricted to zero during FFT. The point could remain marked if it can fit both connectivity and density constraints during following iterations.

When using the real-space filter-2, the output of the filter $f_n(x, y, z)$ is on longer an estimate of the density map; it is instead the input function consisting of the estimate of the map plus some disturbances used to drive the iterating process. The estimate of the density map can be obtained by applying the real-space filter-0 and symmetric operations to the output function from step 4 described above.

The filter-2 is much complicated than filter-1 because there are parameters which need to be specified in order to use the real-space filter-2. These parameters are: the number of largest clusters that should be retained during the iteration (M, and the number of marked points need to be added after the small clusters have been unmarked. The smaller the value of M, the stronger the connectivity constraint is being imposed by the filter.

4.3 Real-space filter-3

The disadvantage of real-space filter-2 is that the connectivity constraint imposed by the filter is less flexible. The constraint is often either too loose or too tight. Under certain circumstance, the real-space filter-3 specified below will provide an alternative in which the connectivity of a map is graduately maximized. However, the real-space filter-3 seems having superior performance only in the cases where there is only one molecule in the asymmetric unit.

The real-space filter-3 consists of following steps:

1. find the clusters formed by the marked points in the asymmetric unit and the size of each cluster;

2. assign a number s to each marked point with s equal to the size of the cluster that the marked point belongs;

3. assign a number s to each unmarked point in the neighborhoods of all marked points with s equals to the size of the cluster that could be formed if the unmarked point becomes marked;

4. calculate the value $e = s \times density$ for each grid point that has s values.

5. sort the grid points according to their e values in the neighborhoods of the marked points in the asymmetric unit;

6. mark the first N grid points with the highest e values and unmark the rest grid points in the asymmetric unit;

7. set the density value of unmarked point be zero;

8. if the density value of a marked point is less than zero, set it to zero; and

9. use the symmetric operations to extend above modifications to the whole unit cell.

4.4 Application example of real-space filter-2

The diffraction data of Mannose Binding Protein (MBP) from 4 wavelength MAD experiment (Burling *et al.*, 1996) is used. The data contains both observed $|F|'s$ and phases. Only $|F|'s$ are used in our iterations. The observed phases are compareed with the calculated ones. The average phase error for the reflections with resolution higher than 10 Åcould be reduced from above 80° to lower 70° by the clustering method for initial density map generating. The phase error for reflections of higher resolution were not reduced significantly by this procedure. The following iterations with real-space filter-2 can improve the molecular envelope but can not reduce the phase error further for the reflections with resolution both above or bellow 10Å.

The MBP crystal has unit cell parameters $a = 65.508$, $b = 72.216$, $c = 45.035$, and $\alpha = \beta = \gamma = 90.0$ and belongs to space group $P2_12_12_1$. The resolution of the diffraction data ranges from 1.065 Åto48.512 Å. In our calculation, there are $64\times64\times64$ grid points in the unit cell. We let N equals to the 8.4/in the asymmetric unit and $M = 150$. The initial density map was created by the clustering method. Each step (or cycle) in the calculation consists of 5 iterations with real-space filter-0 and 1 iteration with real-space filter-2. The results are shown in Fig. 2.

4.5 Application example of real-space filter-3

The myoglobin molecule is used here as an application example for real-space filter-3. The unit cell parameters are: $a = 90.38$, $b = 90.38$, $c = 45.34$, $\alpha = \beta = 90.0, \gamma = 120.0$. The reflections were calculated from the know structure with a solvent mask. The resolution of the reflections ranges from 1.31 Åto 78.27 Å. The unit cell consists of $64 \times 64 \times 32$ grid points. The value N is equal to 12.5% of the total number of grid points. The step of the calculation consists of 20 iterations with real-space filter-0 and 1 iteration with real-space filter-3. The initial map was generated by the clustering method with only 50 repeating times (see step 8 of clustering method). The real-space filter-3 can reduce the average phase error of the reflections with resolution above 4Å. The results are shown in Fig. 3.

5 Discussion

Additional information is needed in order to solve the phase problem in macromolecular crystallography. It seems that the needed information can come from the connectivity of the density map. How precise the molecular structure can be determined depends on how much connectivity information can be provided. The algorithm proposed here could produce a better molecular envelope than that can be done by Subbiah's method (Subbia, 1991, 1993) because the real-space filters described above can provide more connectivity information in comparing with the simple condensing requirement in Subbiah's method.

Is there more detailed information about the general connectivity of density map available at the molecular envelope stage? If such information is available, how can we formulate the information into connectivity constraint?

We have tried to combine our approach with the algorithms of histogram matching, sift-and-rotating method, voting method and patching method. However, none of them can significantly improve the results.

Different optimization approaches such as simulated annealing was tried and was not successful.

We have also tried different type of syntheses including α, β , γ' and $2F_o - F_c$ syntheses in reciprocal space. Unlike that in the approach of skeletonization, we found that $2F_o - F_c$ produced very poor results in our situation where the phase error is relatively large. However, it is still possible that some sort of information input from the reciprocal space can help to overcome the stagnation we are currently facing.

References

Baker, D. Krukouski, A. E. and Agard, D. A. (1993). Acto Cryst. D49:186-192.

Bates, R. H. T. (1982). Optik 61:247-262.

Bystrodd, C., Baker, D., Fletterick, R. J. and Agard, D. (1993). Acta Cryst. D49:429-439.

Dainty, J. C and Fienup, J. R. (1987). In *Image Recovery: Theory and Application.* (Academic Press, New York).

Greer, J. (1985). Meth. Enzymol. 115:206-226.

Hauptman, H.A. (1989). Physics Today. 42:24-29.

Hayes, M. H. (1982). IEEE Trans. Acoustics, Speech, and Signal Proceesing 30:140-154.

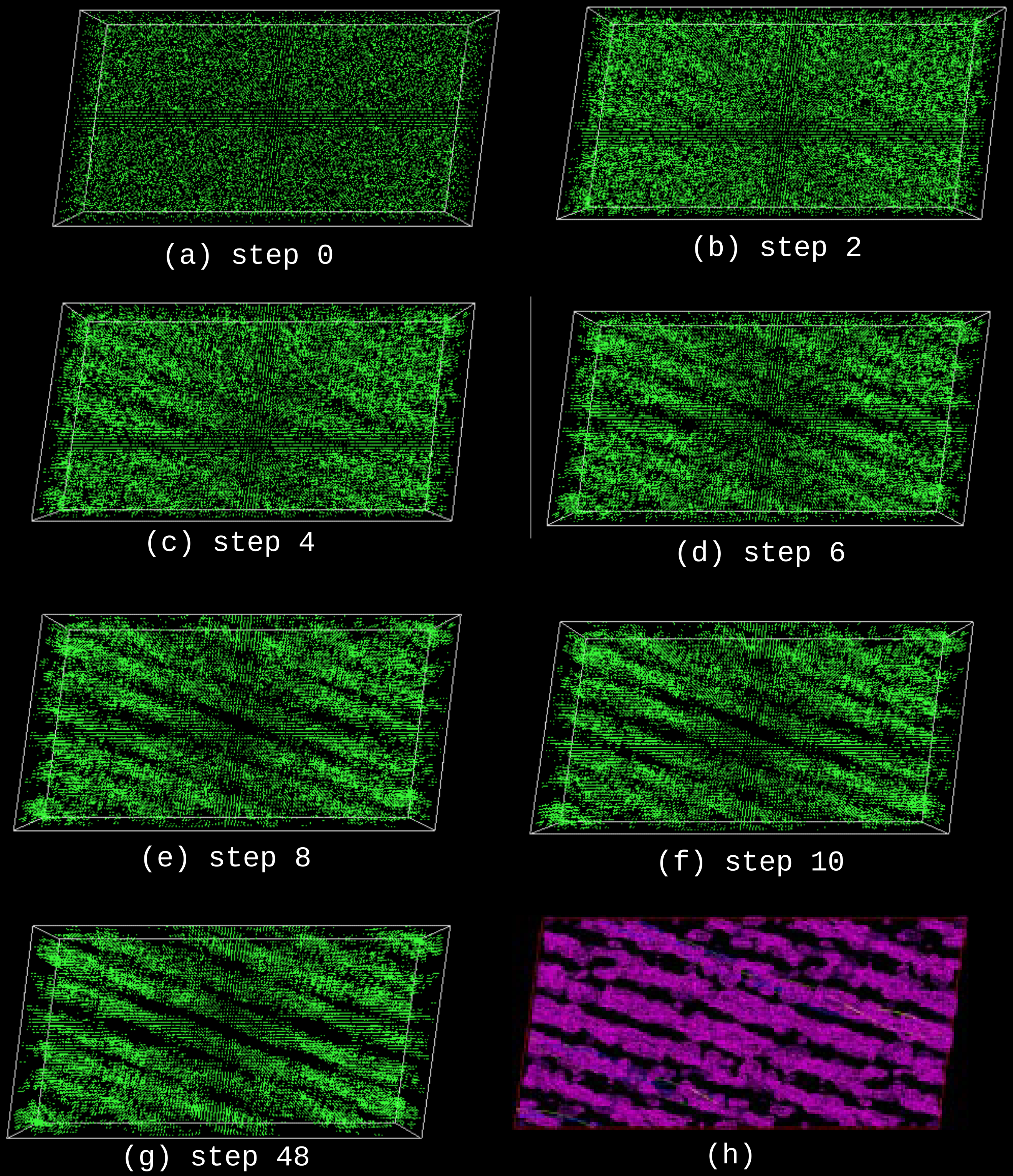
Hayes, M. H. (1987). In *Image Recovery: Theory and Application.* (Academic Press, New York).

Karle, J. and Hauptman, H. (1956). Acta crystallogr. 9:635-651.

Millane, R. P. (1990). J. Opt. Soc. Am. 7:394-411.

Stark, H. (1987). In *Image Recovery: Theory and Application.* (Academic Press, New York).

Wilson, C. and Agard, D. A. (1993). Acto cryst. A49:97-104.



(a) step 0

(b) step 2

(c) step 4

(d) step 6

(e) step 8

(f) step 10

(g) step 48

(h)

Fig. 1a. Applying the iteration algorithm with real-space filter-1 to the Tropomyosin data. (a) to (g) show the first 50 iterations. (h) is the SIR-Fo map from 7A model.

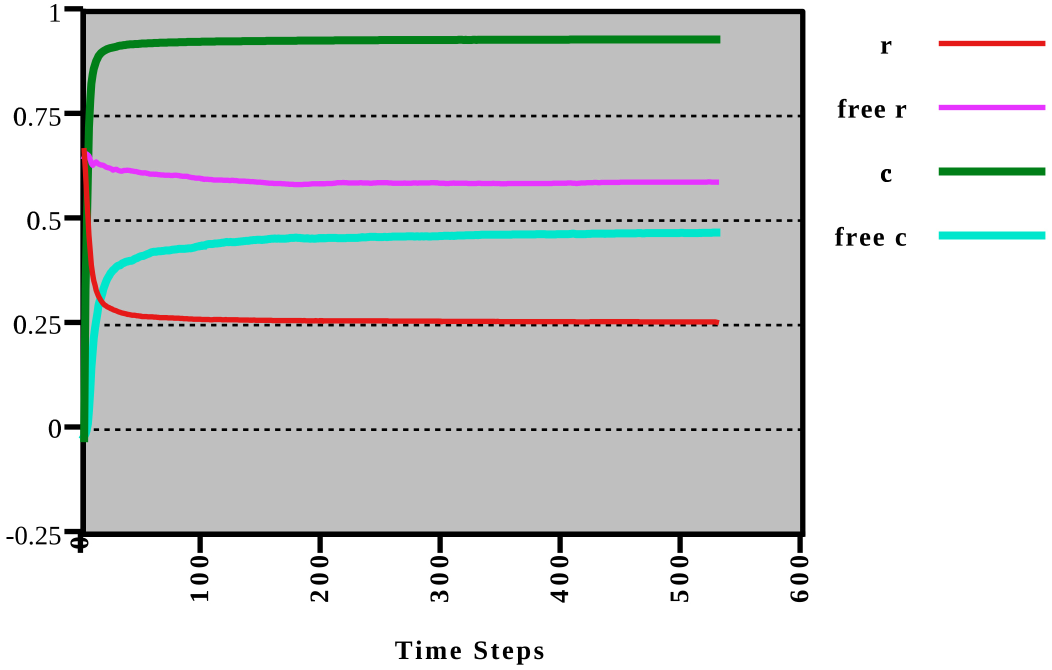


Fig. 1b. The changes of r, free−r, correlation coefficient (c), and free−c betwee calculated and measured |F|’s of tropomyosin at each step of iteration.

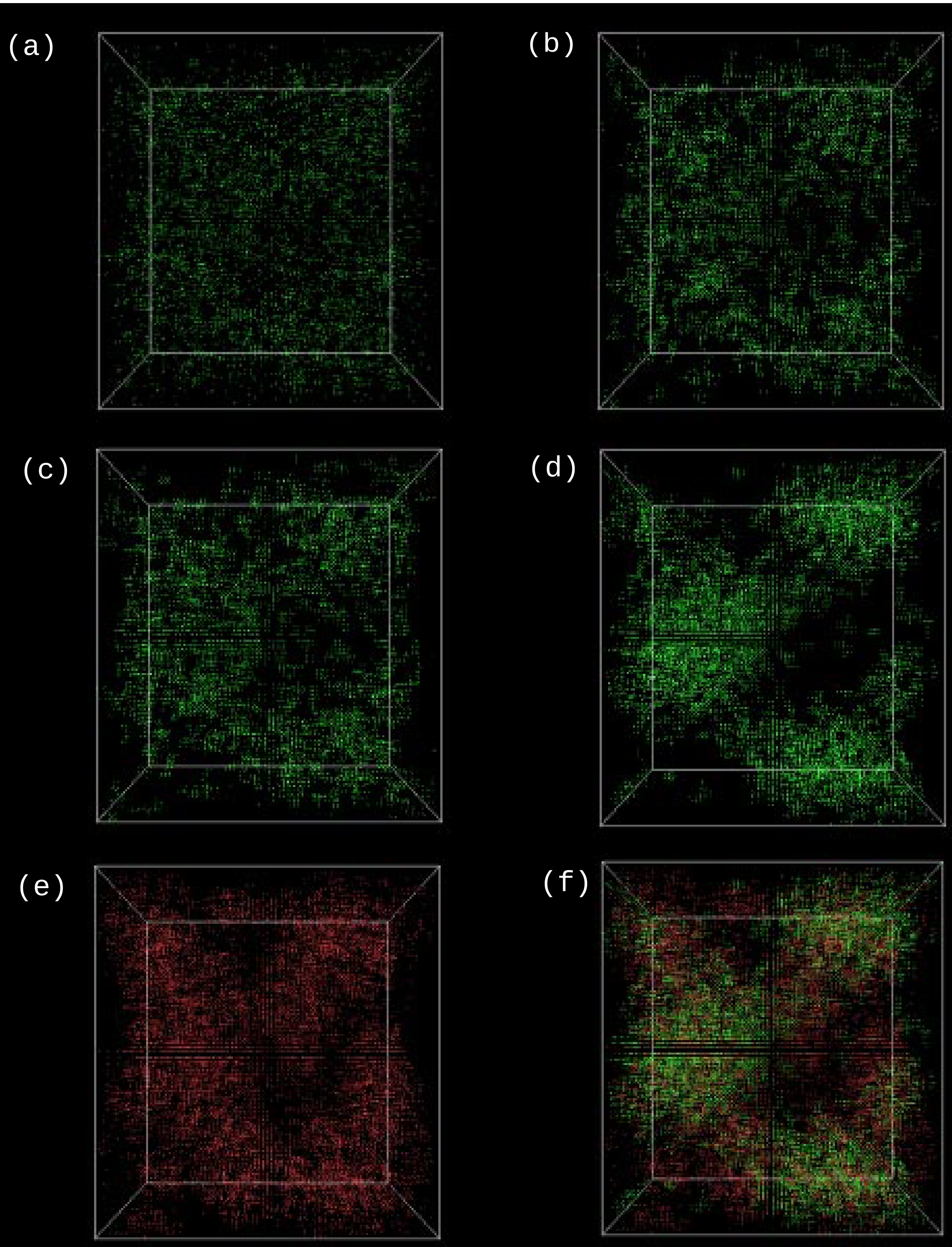


Fig. 2. Iterations with the real−space filter−2 of MBP data. (a) is the initial map generated by the clustering method. (b), (c) and (d) are step 1, 2 and 150 of the interation, respectively. (e) is the map calculated from the measured phases. (f) is the overlap of (d) and (e).

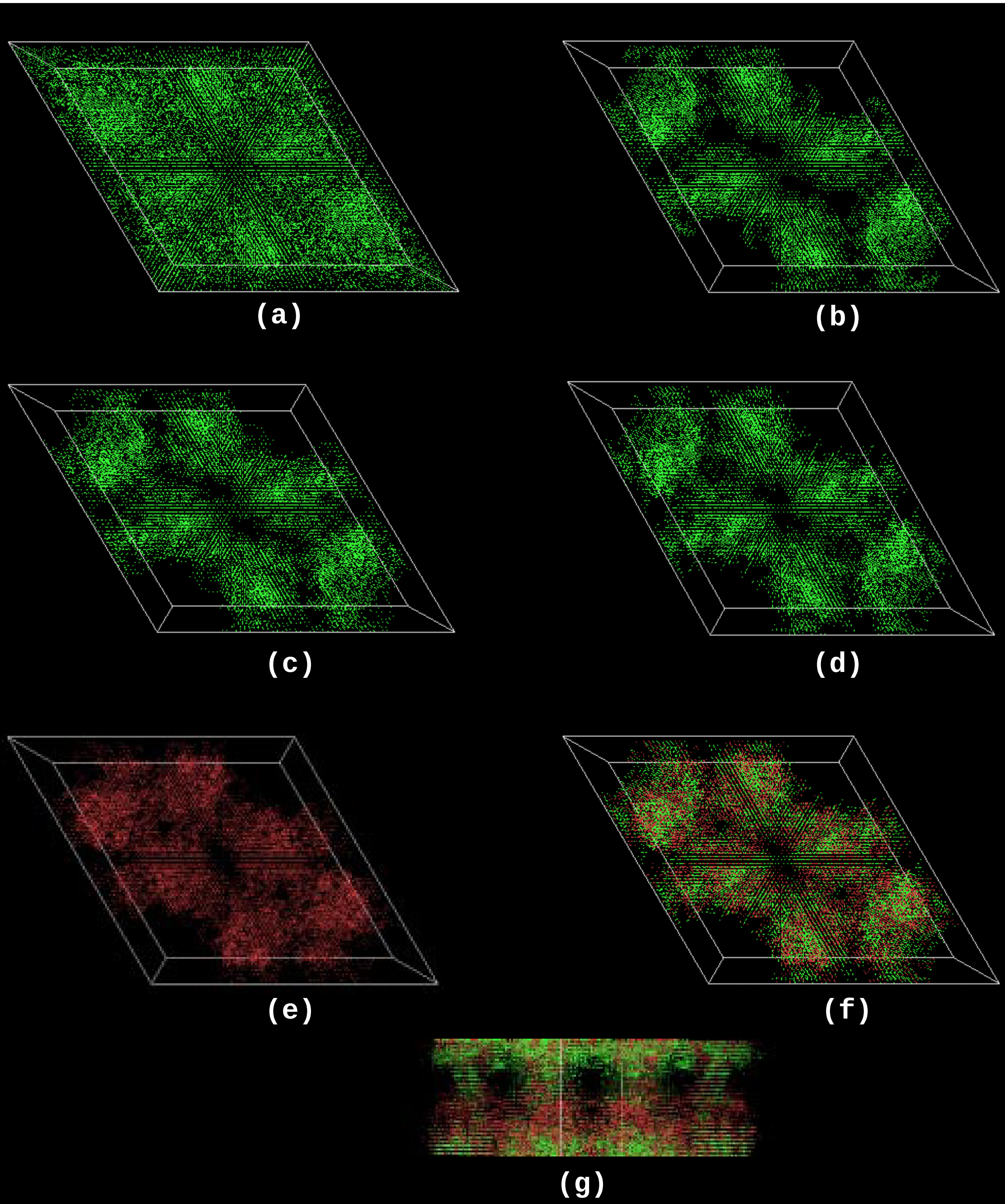


Fig. 3a. Iterations with the real-space filter-3 of myoglobin data. (a) is the initial map generated by the clustering method. (b), (c) and (d) are step 1, 2 and 200 of the interation, respectively. (e) is the map calculated from the known structure. (f) and (g) show the overlap of of (d) and (e) on x-y and x-z plans, respectively.

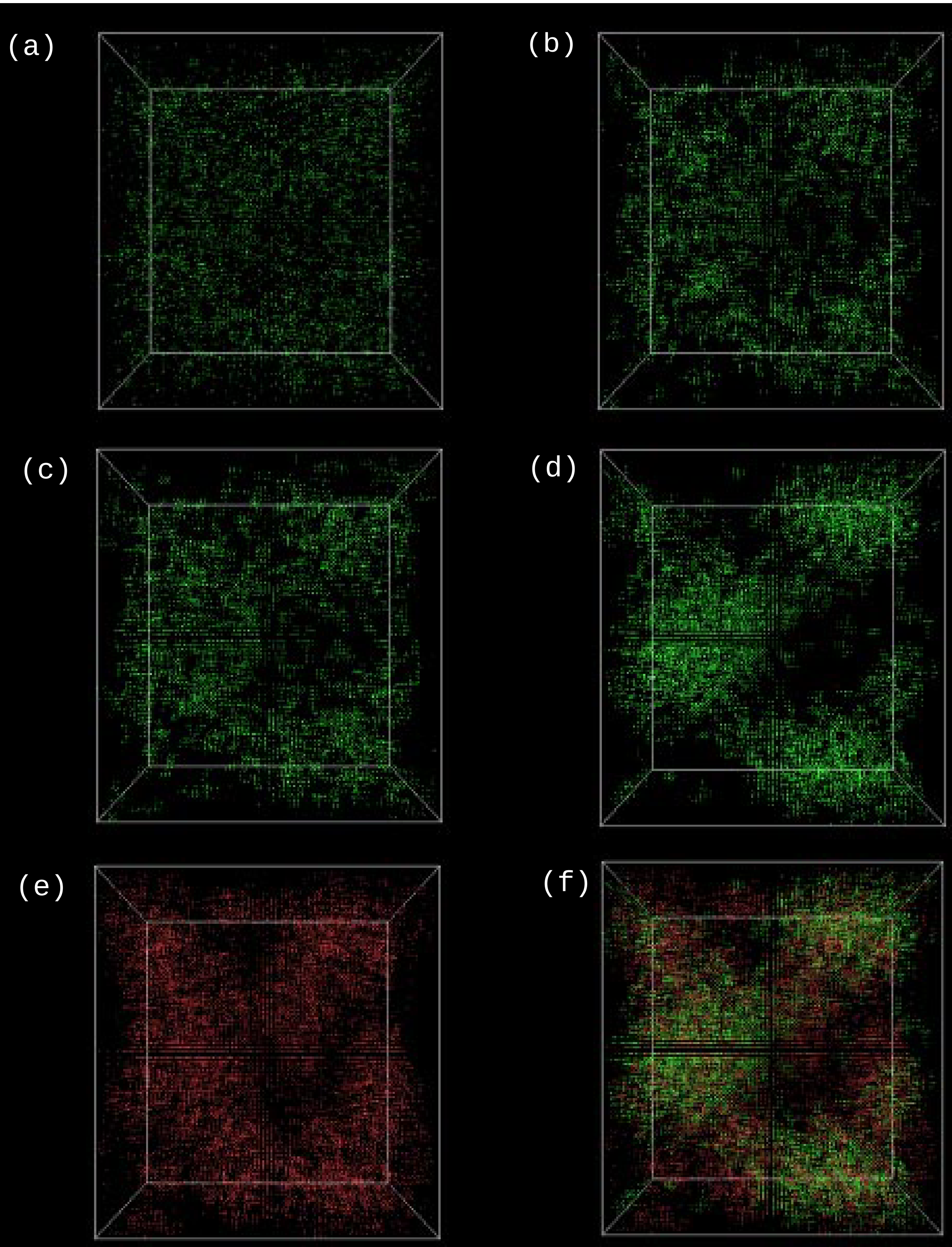


Fig. 2. Iterations with the real−space filter−2 of MBP data. (a) is the initial map generated by the clustering method. (b), (c) and (d) are step 1, 2 and 150 of the interation, respectively. (e) is the map calculated from the measured phases. (f) is the overlap of (d) and (e).

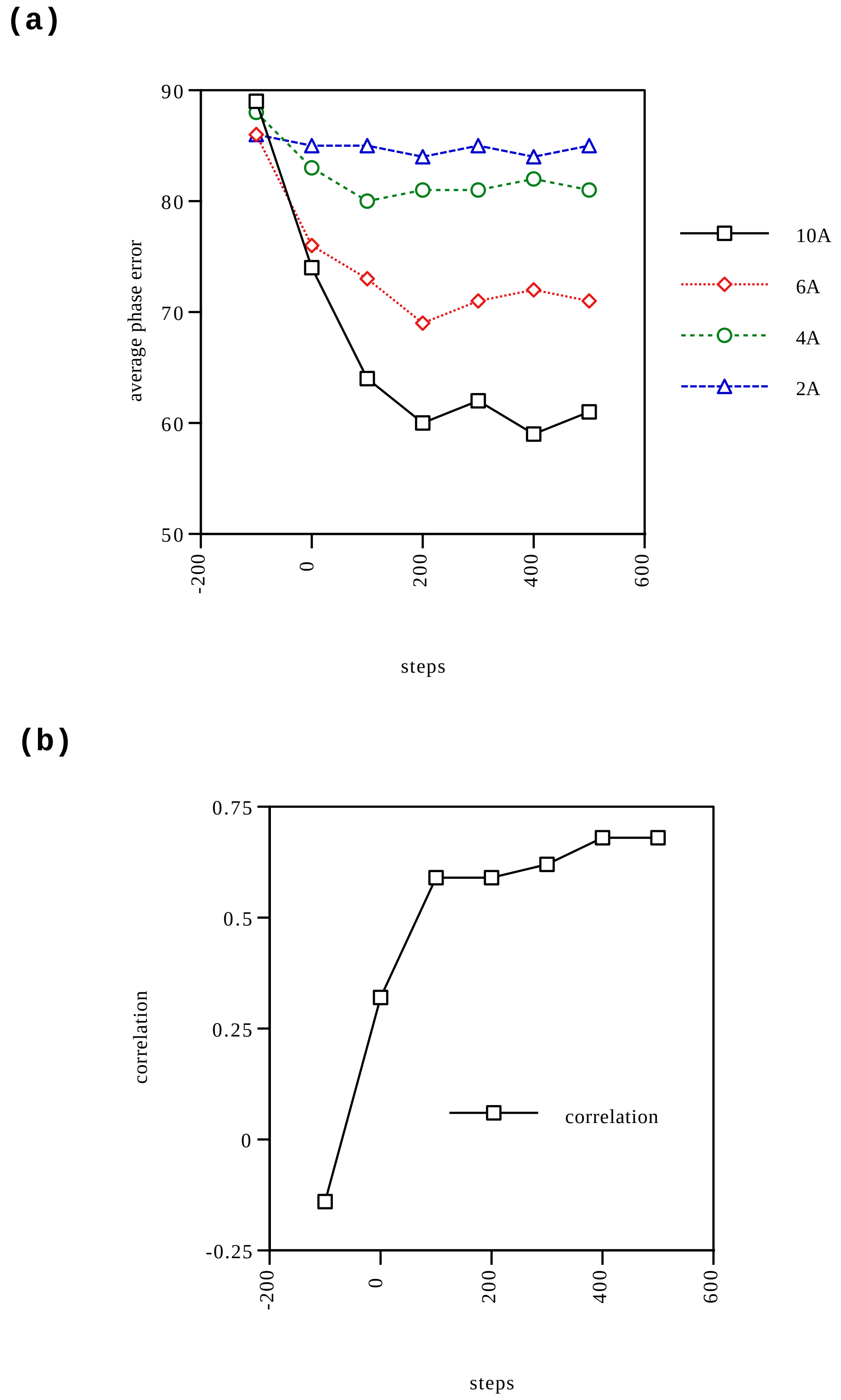


Fig. 3b. Change of the average phase errors at different resolution (a), and the correlation coefficient during the iterations of myoglobin data.