A new implementation of the molecular replacement method using a six-dimensional Patterson vector search

Fan Jiang[a,b] and Zihe Rao[a,b]*

[a]Laboratory of Structural Biology, Department of Biological Sciences and Biotechnology, Tsinghua University, Beijing 100084, People's Republic of China, and [b]Protein Science Laboratory of MOE, Tsinghua University, Beijing 100084, People's Republic of China. E-mail: raozh@xtal.tsinghua.edu.cn

The current molecular replacement programs are primarily implemented in reciprocal space. In this paper a new implementation in direct (real) space is proposed by matching the model atomic vectors with the vectors in the Patterson vector space using a six-dimensional exhaustive search method. It is shown that this implementation can find the correct rotations and translations of α helices in a myoglobin crystal structure using experimental diffraction data at 2 Å resolution. A comparison with previous Patterson vector search methods is discussed.

Keywords: molecular replacement; Patterson vector search; six-dimensional search.

1. Introduction

The molecular replacement method (Rossmann & Blow, 1962) is a very powerful and efficient method of solving the phase problem when part of the target unknown structure is known. As implemented in reciprocal space, it consists of two steps: rotation search and translation search. In rotation search, the Patterson vectors (map) of the search model are matched with those of the target crystal. An integration radius is chosen so that only the self-Patterson vectors are matched in rotation search. In translation search, a Patterson correlation function is calculated. There are many implementations of the rotation search and the translation search, of which *AMoRe* (Navaza, 1987) and *X-PLOR* (Brunger *et al.*, 1987; Huber, 1965) are the most popular program packages and are widely used. Because the overlap of the self-Patterson vectors is very serious, when the search model is only a small part of the target structure the rotation search often fails to produce the correct solutions. Experience shows that the search model should not be less than a quarter of the target structure content. When the rotation solutions are inaccurate, it is impossible for the translation search to find the correct solutions. This is the main limitation in applying the molecular replacement method.

Historically, the molecular replacement method has also been implemented in real space (Hoppe & Paulus, 1967; Nordman & Nakatsu, 1963; Nordman, 1966; Schilling, 1970). Furthermore, the real-space implementation has been recently applied to the solution of macromolecular structures (Nordman, 1972, 1994). With the advent of more powerful computers, it is possible to re-implement the molecular replacement method in real space with more efficient algorithms. In this paper, we provide the formula with which we implement an algorithm for calculating all the interatomic vectors between two symmetry-related search models and matching them with the cross-Patterson vectors. A fast translation algorithm is implemented as developed previously (Jiang & Kim, 1991) so that an exhaustive rotation search can be achieved. We show that using the 2 Å experimental diffraction data the correct rotations and translations can be found for all the α helices in myoglobin using only the main-chain atoms in the search model. We discuss our results in comparison with previous implementations and suggest the directions of future developments.

2. Methods

2.1. *Derivation of the matching formula*

We denote x_i and x_j as the atomic vectors of the search model; v_k as the Patterson vectors of the target structure; R and t as the rotation matrix and translation vector of the rigid-body transformation applied to the search model; S_1, t_1 and S_2, t_2 as the two different symmetry operations,

$$\overrightarrow{x_i'} = R\overrightarrow{x_i} + \overrightarrow{t}, \qquad (1)$$

$$\overrightarrow{x_j'} = R\overrightarrow{x_j} + \overrightarrow{t}, \qquad (2)$$

$$\overrightarrow{x_i''} = S_1\overrightarrow{x_i'} + \overrightarrow{t_1}, \qquad (3)$$

$$\overrightarrow{x_j''} = S_2\overrightarrow{x_j'} + \overrightarrow{t_2}, \qquad (4)$$

$$\overrightarrow{v_k} = \overrightarrow{x_j''} - \overrightarrow{x_i''} = S_2R\overrightarrow{x_j} + S_2\overrightarrow{t} + \overrightarrow{t_2} - S_1R\overrightarrow{x_i} - S_1\overrightarrow{t} - \overrightarrow{t_1}, \qquad (5)$$

$$\overrightarrow{v_k} + \overrightarrow{t_1} - \overrightarrow{t_2} = S_2R\overrightarrow{x_j} - S_1R\overrightarrow{x_i} + (S_2 - S_1)\overrightarrow{t}. \qquad (6)$$

2.2. *Implementation*

Our translation search algorithm is very fast and has been developed previously for docking two molecular surfaces (Jiang & Kim, 1991). Briefly, all difference vectors between two sets of vectors, namely, the search and the target vectors, are calculated and the matching score between each pair of the search and target vectors is accumulated in a translation vector matrix. After looping through all different pairs of the search and target vectors, the translation vectors with the highest matching scores are found from the translation vector matrix. Our rotation search is exhaustive. The rotation space is sampled with polar angles (ϕ, φ, χ) and the polar angles are sampled with grids.

After the rotation and translation search, all R and t are sorted in descending order of the matching scores and the sorted solutions are used for clustering. The clustering algorithm is simple. A rotation distance cut-off (in degrees) and a translation distance cut-off (Euler distance) are selected. The clusters are searched from the top-score solutions downward. The first solution is a new cluster. Then, if the next solution is outside the range of the rotation and translation distance cut-offs, a new cluster is generated and saved. In this way, similar (neighboring) solutions are grouped together and the uneven sampling in the rotation space is also removed. In our tests, the rotation distance cut-off and the translation distance cut-off are 25° and 10 Å, respectively. The choices of the relatively big cut-offs take into account the fact that the errors of the solutions can be relatively large and that the small cut-offs will diminish the purpose of clustering the solutions. It is also noted that the α helix has self-symmetry, *i.e.* there are multiple ways of superimposing a helix onto itself. The solutions related by the helix self-symmetry operations are grouped together. In our tests, the known correct solutions are compared with the clustered solutions.

Table 1
Results of the Patterson vector search.

The search model is the main-chain atoms of residues 3 to 18 from the structure 104M. Column 2 shows the residues of the individual α helices in the target structure 1A6M. Column 3 shows the root-mean-square deviation between the main-chain atoms of the search model and the individual target helices.

Helix number	Residues	Root-mean-square deviation (Å)	Solution rank	Correlation coefficient
1	3–18	0.11	1	0.799
2	20–35	0.76	2	0.782
3	36–42	0.65	3	0.772
4	51–57	0.60	8	0.775
5	58–77	0.62	7	0.775
6	86–94	0.56	4	0.781
7	100–118	0.65	6	0.776
8	124–149	0.85	5	0.779

The 'image-seeking function' we selected for the current implementation is the correlation coefficient between the Patterson vector peak heights and the interatomic vector weights, as suggested by Nordman (1994). It has been pointed out that the interatomic vectors are not always located at the peak position in the Patterson map (Buerger, 1959). Therefore, we do not use the point atoms to calculate the interatomic vectors of the model. Instead, we first calculate a model electron density map from the search model at a proper resolution and select all the density grid points above a certain peak height (*e.g.* 2σ), and then calculate the model interatomic vectors from these selected grid points. *CCP4* programs were used in these calculations (Collaborative Computational Project, Number 4, 1994).

3. Results and discussion

We used an α helix (residues 3 to 18) of a myoglobin crystal structure (PDB code 104M) as our search model. The structure 104M is sperm whale myoglobin, belonging to space group $P2_1$ with cell dimensions of $a = 64.73$, $b = 30.91$, $c = 34.83$ Å and $\beta = 105.41^\circ$. The experimental diffraction data used in our tests was retrieved from RCSB (www.rcsb.org) with a PDB code 1A6M, also a sperm whale myoglobin structure, belonging to space group $P2_1$ with cell dimensions of $a = 63.80$, $b = 30.81$, $c = 34.35$ Å and $\beta = 105.80^\circ$. Only reflections up to 2 Å were included in our calculations.

The results are shown in Table 1. It can be seen that all the helices in myoglobin could be located in the top eight clustered solutions with the highest correlation coefficients. These results are similar to those of a previous study (Nordman, 1972) in which individual helices were also searched in Patterson vector space and the correct orientations and translations were found. The difference between our current implementation and that of Nordman (1972) is that the latter used a two-stage search strategy: use the intramolecular (self) vectors to find the rotation and then use the intermolecular (cross) vectors to find the translation. In our implementation we utilize the fact that the rotation information is not only contained in the self-Patterson vectors but also in the cross-Patterson vectors. A six-dimensional search (in $P2_1$, a five-dimensional search) can find the rotation and the translation of a search model simultaneously. It is not surprising that similar results have been obtained. Our implementation is computationally more intensive but reachable with the current computing power (2 h on Intel Pentium III 450 Hz). We believe the six-dimensional search method will prove to be more sensitive and useful in future developments. This is because the six-dimensional search method avoids the crowding of the self-Patterson vectors in the rotation search stage, which has two ramifications. One is an increased tolerance of the errors in the search model and thus a larger radius of convergence than the two-stage search method. The other is that even smaller known structures than those used in our present tests could be used as a search model in molecular replacement. Further testing is needed to demonstrate these advantages of our approach. A few other image-seeking functions have been suggested previously (Nordman, 1994) and shown to be effective. In our implementation the correlation coefficient is more easily implemented and requires the least amount of computation. We will try to implement other image-seeking functions in the future with more efficient algorithms.

Although the two crystal structures used in our tests are very similar, as suggested by their cell parameters and space groups, our tests were not performed on an ideal case but instead used two experimental structures with the reflection data of one of them available, both structures determined and deposited independently as entries 104M and 1A6M, respectively. Since a single search model, consisting of the main-chain atoms of residues 3 to 18 from the structure 104M, was used, the errors between the search model and the target fragment were not as small as the overall difference between the two structures, 104M and 1A6M, might have suggested. The root-mean-square deviation between the main-chain atoms of the two structures is 0.24 Å, while those between the search model and the individual target helices are listed in Table 1. These listed root-mean-square deviations should be comparable with the value one might expect for the difference between an ideal helix and a regular α helix in any globular protein. Therefore, the overall similarity between the structures 104M and 1A6M should not affect the generality of our test results.

It is worth noting that using the same search model, *i.e.* a helix consisting of residues 3 to 18 of the structure 104M, we could not find the correct rotations with other available reciprocal-space implementations such as *AMoRe*, *X-PLOR* and *CNS* (data not shown). Because we use grid points to represent the Patterson map and the search model in the form of a calculated electron density map, we can choose different resolution ranges for map calculations so that different levels of details of the search model can be included and different amounts of diffraction data can be selected. More tests will be performed using different resolution ranges.

Recently, several algorithms have been developed for performing six-dimensional searches in molecular replacement (Kissinger *et al.*, 1999; Chang & Lewis, 1997; Tong, 1996). However, they are all implemented in reciprocal space. Among them, the method of Kissinger and co-workers (Kissinger *et al.*, 1999) is the latest, implemented in program *EPMR*, and has been tested extensively on a variety of structures. We will discuss the relevant differences between *EPMR* and our method in the following.

First, Kissinger *et al.* (1999) have shown that *EPMR* is very efficient and fast as the number of required structure-factor calculations to achieve the six-dimensional search is considerably less than that if a systematic six-dimensional search is conducted. In fact, according to their estimation, a systematic six-dimensional search in reciprocal space would have been computationally infeasible. In contrast, we have shown that a systematic six-dimensional search is possible when conducted in real space using our proposed algorithm. Second, Kissinger *et al.* (1999) demonstrated that *EPMR* could use less accurate or less complete search models. In the test case of 6RHN, the error for polyalanine atoms was 0.30 Å and the maximum truncation achieved was 60%. In our test case of myoglobin (1A6M) the average error between the helices was ~0.7 Å and the truncation used was almost 90% (using only 16 residues out of 153 residues in myoglobin). Third, *EPMR* has been tested on a variety of structures and shown to be able to tolerate errors as large as 3 Å (without truncation), better than *CNS* and *AMoRe*. Although *CNS* and *AMoRe* could not produce correct solutions in our test case, we have not tested our method on search models with such large errors. The first two differences represent significant advantages of our method while the third difference points to one of the directions of our future development. We would also like to point out that our intended development of this systematic six-dimensional search method in real space is not only for conventional molecular replacement using large search models, but, more importantly, for the purpose of using increasingly smaller fragments such as helices and sheets as search models, with the hope that this approach will eventually solve the phase problem for macromolecules. Therefore, our present work should not be viewed solely from the perspective of rivaling the currently available molecular replacement methods for conventional structure determination. We believe that our preliminary results are encouraging and the further pursuit of our method is warranted.

In summary, we have presented here a new implementation of the molecular replacement method in real space using a six-dimensional exhaustive search of Patterson vector space. When a search model consisting of an α helix from residues 3 to 18 from a myoglobin structure (104M) was used, all other helices in another myoglobin structure (1A6M) could be found, using the 2 Å experimental data for 1A6M which was available. Our results are similar to those of a previous study using a two-stage vector search method in real space. We believe our current implementation deserves further development and testing in order to fully explore its potential applications.

We thank the support of Tsinghua University Research Grant 985, National Basic Research Fund 973 (Grant Nos. G1999075602, G1999011902, G1998051105), and National Science Foundation of China (Grant Nos. 39870174, 39970155).

References

Brunger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458–460.
Buerger, M. J. (1959). *Vector Space*. New York: John Wiley.
Chang, G. & Lewis, M. (1997). *Acta Cryst.* D**53**, 279–289.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Hoppe, W. & Paulus, E. F. (1967). *Acta Cryst.* **23**, 339–342.
Huber, R. (1965). *Acta Cryst.* **19**, 353–356.
Jiang, F. & Kim, S.-H. (1991). *J. Mol. Biol.* **219**, 79–102.
Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst.* D**55**, 484–491.
Navaza, J. (1987). *Acta Cryst.* A**43**, 645–653.
Nordman, C. E. (1966). *Trans. Am. Crystallogr. Assoc.* **2**, 29–38.
Nordman, C. E. (1972). *Acta Cryst.* A**28**, 134–143.
Nordman, C. E. (1994). *Acta Cryst.* A**50**, 68–72.
Nordman, C. E. & Nakatsu, K. (1963). *J. Am. Chem. Soc.* **85**, 353–354.
Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.
Schilling, J. W. (1970). *Crystallographic Computing*, edited by F. R. Ahmed, p. 115. Copenhagen: Munksgaard.
Tong, L. (1996). *Acta Cryst.* A**52**, 782–784.

Received 27 September 2000 • Accepted 9 January 2001