Improvement of a new rotation function for molecular replacement by designing new scoring functions and dynamic correlation coefficient*

Jiang Fan(江 凡)[a)†] and Ding Wei(丁 玮)[a)b)]

[a)] *Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing* 100190, *China*

[b)] *Graduate School of the Chinese Academy of Sciences, Beijing* 100049, *China*

(Received 26 April 2010; revised manuscript received 10 May 2010)

A previously published new rotation function has been improved by using a dynamic correlation coefficient as well as two new scoring functions of relative entropy and mean-square-residues to make the rotation function more robust and independent of a specific set of weights for scoring and ranking. The previously described new rotation function calculates the rotation function of molecular replacement by matching the search model directly with the Patterson vector map. The signal-to-noise ratio for the correct match was increased by averaging all the matching peaks. Several matching scores were employed to evaluate the goodness of matching. These matching scores were then combined into a single total score by optimizing a set of weights using the linear regression method. It was found that there exists an optimal set of weights that can be applied to the global rotation search and the correct solution can be ranked in the top 100 or less. However, this set of optimal weights in general is dependent on the search models and the crystal structures with different space groups and cell parameters. In this work, we try to solve this problem by designing a dynamic correlation coefficient. It is shown that the dynamic correlation coefficient works for a variety of space groups and cell parameters in the global search of rotation function. We also introduce two new matching scores: relative entropy and mean-square-residues. Last but not least, we discussed a valid method for the optimization of the adjustable parameters for matching vectors.

Keywords: molecular replacement, rotation function, global search, correlation coefficient

PACC: 6110M, 6150K, 8710, 8715

1. Introduction

As structural genomics progresses, automatic structure determination using the molecular replacement method has been becoming increasingly important.[1] This has spurred more development of both the search engine and the automatic intelligent pipeline for molecular replacement such as PHASER, Mr.Bump, AMORE, ARP/wARP, the JCSG pipeline and MOLREP.[2−7]

The search engine for molecular replacement method is primarily concerned with a problem that when a search model is given, how the search engine (a program or a set of programs) finds the best orientation and position of this model in the target crystal.[8] There are six degrees of freedom, three for orientation (the rotation search) and three for position (the translation search). According to the manner in which these six degrees of freedom are searched, the search engine can be divided into three types: one is to perform the rotation search first and then the translation search; the second is to perform the rotation and translation search simultaneously such as in EPMR.[9] and Queen-of-Spades;[10] the third is to do a brute-force sampling of the rotation space and for each rotation sampled perform a translation search. For any given translation search method that is sufficiently fast, the third type search engine (or search strategy) may be used. For a structure determination problem using molecular replacement, a combination of the above three search strategies may be used for either finding the initial solutions or refining potential good solutions.

Previously we have introduced a new rotation function for the rotation search of the molecular replacement method.[11] In this new method, the atomic coordinates of the search model is directly matched

*Project supported by the National Natural Science Foundation of China (Grant Nos. 10674172 and 10874229).

†Corresponding author. E-mail: fjiang@aphy.iphy.ac.cn

© 2010 Chinese Physical Society and IOP Publishing Ltd

with the Patterson map of the target crystal. The goodness of matching is measured by several scoring functions such as the product of atomic and map weights, the correlation coefficient of atomic and map weights, the relative R-factor of atomic and map weights, and so on. The signal-to-noise ratio of the matching is enhanced by first averaging the peaks of all possible matches and then optimizing a set of weights by the linear regression method to obtain a combined score of all the individual matching functions. We have found that the set of weights optimized from local search around the correct solution can be generalized to global search with reasonably good results. This set of weights should at least be applicable for a given model and a target crystal with known space group and cell parameters.[11] We also found the set of weights optimized depends on the crystal form. This will greatly limit the applicability of our new rotation function when applying to an unknown target crystal form.

In this work, we propose a new method of calculating correlation coefficient for each rotation angle sampled in the rotation function search, and this "dynamic" correlation coefficient is used as the value of our new rotation function. So the set of weights are optimized dynamically and locally instead of globally for a given target crystal, this will reduce the dependence of the weights on the crystal form. In addition, we add two new matching scoring functions: the relative entropy between the model weights and the Patterson map weights, and the mean-square residues between the model weights and the map weights. The relative scales between the model and map weights are such to minimize their differences in distribution. We found that our new rotation function was applicable to a number of crystal structures extracted from PDB with experimental diffraction data.

2. Materials and methods

2.1. Relative entropy as a matching function

The relative entropy between the weights of the search model and Patterson map is calculated by the following formula:

$$R_{\rm ent} = \frac{1}{N} \sum_{i}^{\{MP\}} p_i \ln\left(\frac{p_i}{q_i}\right) - \langle p\rangle \ln\left(\frac{\langle p\rangle}{\langle q\rangle}\right). \quad (1)$$

Here the summation is over the set of the matched points between the model and Patterson map, $\{MP\}$, whose total number is N. p_i is the weight of the Patterson map vectors and q_i is the weight of the atomic vectors of the search model; $\langle p\rangle$ and $\langle q\rangle$ are the averaged weights over the matched points (or vectors).

The initial relative scale is calculated by normalizing the weights over the whole search model and Patterson map vectors. Then for each translation vector in the rotation search, the weight distribution is renormalized for each set of matched points by subtracting the second term in the above equation.

2.2. Mean-square residues as a matching function

The mean-square residues between the weights of the search model and Patterson map are calculated by the following formula:

$$R_{\rm res} = \frac{1}{N} \sum^{\{MP\}} (p_i - q_i)^2. \quad (2)$$

Here the summation is over the set of the matched points between the model and Patterson map, $\{MP\}$, whose total number is N. p_i is the weight of the Patterson map vectors and q_i is the weight of the atomic vectors of the search model. The initial relative scale is calculated by normalizing the weights over the whole search model and Patterson map vectors. However, in this case the weights are not renormalized for the matched points.

2.3. Dynamic correlation coefficient as the rotation function

The dynamic correlation coefficient is calculated as follows. First, all the globally sampled rotations are applied to the search model and its matching scoring functions are calculated. Secondly, for each rotation angle, its closest neighbouring angles sampled are found. This is done for each rotation angle sampled only once and the indexed neighbour list can be reused for different calculations of dynamic correlation coefficients. From the neighbour list, a given number of closest angles are selected randomly and the set of angles and their matching scores are stored in a file or matrix. Thirdly, this set of matching scores are submitted to a linear regression fit using the angular distance from each neighbouring angle to the

rotation angle used to define the neighbour list, as the fitting "experimental" value. Finally, after the linear regression fitting, there are a calculated or fitted value and an "experimental" value. The correlation coefficient between these two sets of values is defined as the dynamic correlation coefficient for the given rotation angle. Since the dynamic correlation coefficient is calculated by using its local angle neighbours, it can also be called the local dynamic correlation coefficient. This value should better reflect the goodness-of-fit between the search model and the Patterson map at this particular rotation angle.

2.4. Adjustable parameters vector matching

Rotation angle sampling is described previously.[11] Grid size is the grid spacing used for digitizing the translation space. Nsum_cut-off_unit is the number of matched vectors (or points) in calculating the matching scores. Nsum_cut-off_asymmunit is the number of the averaged matched vectors after all translation vectors have been folded into an input asymmetric unit determined by the crystal space group. Both Nsum values are cut-off to set the minimum required for including the corresponding scores in the overall average over all possible translation vector peaks.

2.5. Test cases

The search model was the same as in the previous publication. In this work, we only used a poly-Ala model. It consists of a helix-turn-helix from residues 3 to 36 in the myoglobin structure with PDB code 1a6m. The nine target crystals used in this work are listed in Table 1. Both the coordinates and diffraction data were extracted from the PDB (www.rcsb.org). All of them have experimental diffraction data available for download up to 2.0 Å (1 Å = 0.1 nm) except PDB entry 2mm1. The Patterson maps for all tests were calculated using CCP4 programs and the resolution cut-offs for the diffraction data were all 2.0 Å. The Patterson vectors selected for matching were map points with values greater than or equal to 1σ. One unit cell of the Patterson map was scanned for matching.

3. Results

The target crystals cover a range of space groups and cell parameters, as listed in Table 1. Tests were performed using these crystal structures, where global searches were conducted using polar angles and the rotation solutions were ranked using dynamic correlation coefficients. The search model was the poly-Ala model extracted from the structure of 1a6m from residues 3 to 36, which is a unique fragment within the myoglobin. The nine crystal structures were used as the "unknown" target structures. The main adjustable parameters are the grid size, the minimal number of matched vectors (Nsum_cut-off_unit), and the minimal number of matched vectors within an asymmetric unit (Nsum_cut-off_asymmunit). After screening a range of parameters for grid size from 4.6 to 5.4, Nsum_cut-off_unit from 2 to 16, and Nsum_cut-off_asymmunit from 2 to 16, the best results are listed in Table 2. Choice of the asymmetric unit for each crystal is the smallest possible for the given space group in the fractional coordinates. Other choices of the same size give similar results. From Table 2 it can be seen that our new rotation function using dynamic correlation coefficient has found good correct solutions for all nine target crystals, as well as for the crystal of the search model. These crystals cover a range of space groups and cell parameters, indicating that the new improved method of calculating the rotation function is not limited to a particular space group and does not need to be optimized relative to a specific search model.

Criteria for good solutions are that the angular distances of the solutions from the correct target structure are not greater than 15 degrees and the ranks of the solutions are not greater than 1000. After the rotation search, the translation search is performed around the rotation angle suggested by sampling its neighbouring angles, which constitutes a local refinement of the orientation. So as long as there is at least a good solution obtained by the rotation search, it is possible to find the target solution in the translation search. Of course, it is generally better to have more number of good solutions and secondarily better to have less ranking number for good solutions to increase the chance of success in the translation search. The results in Table 2 suggest that a big step has been made toward the practical application of this new rotation function.

Table 1. Crystals used for test cases.

crystal	space group	cell parameters	resolution/Å	reference
1a6m	$P2_1$	$a = 63.800$, $b = 30.810$, $c = 34.350$, $\beta = 105.80°$	1.78	[12]
1dti	$P2_12_12_1$	$a = 49.163$, $b = 40.002$, $c = 80.011$	1.70	[13]
102m	$P6$	$a = b = 91.433$, $c = 45.949$, $\gamma = 120.00°$	1.84	
1j3f	$P2_12_12_1$	$a = 32.972$, $b = 58.787$, $c = 76.237$	1.45	[14]
1mlu	$P6$	$a = b = 91.200$, $c = 45.870$, $\gamma = 120.00°$	1.90	[15]
1wvp	$P4_32_12$	$a = b = 60.982$, $c = 84.808$	1.53	[16]
2evp	$P2_12_12_1$	$a = 39.670$, $b = 47.770$, $c = 77.430$	1.70	[17]
2jho	$P2_1$	$a = 34.409$, $b = 29.786$, $c = 63.633$, $\beta = 104.93°$	1.40	[18]
2mm1	$P3_221$	$a = b = 86.200$, $c = 35.600$, $\gamma = 120.00°$	2.80	[19]

Table 2. Number of good solutions and best ranking of correct solutions for global searches using dynamic correlation coefficient.

crystal	space group	asymmetric unit	grid size/Å	Nsum cut-off for matching[a]	Nsum cut-off for asymmetric unit[b]	number of good solutions[c]	best ranking of correct solutions[d]
1a6m	$P2_1$	1.0 1.0 0.5	5.2	4	4	8	139
1dti	$P2_12_12_1$	1.0 0.5 0.5	4.8	4	4	44	42
102m	$P6$	0.5 0.5 1.0	5.0	2	12	1	2
1j3f	$P2_12_12_1$	1.0 0.5 0.5	5.2	2	12	3	207
1mlu	$P6$	0.5 0.5 1.0	5.4	2	8	4	105
1wvp	$P4_32_12$	0.5 0.5 1.0	5.4	2	2	4	490
2evp	$P2_12_12_1$	1.0 0.5 0.5	5.0	4	4	27	6
2jho	$P2_1$	1.0 1.0 0.5	5.0	4	8	5	3
2mm1	$P3_221$	1.0 1.0 0.5	5.0	2	8	3	316

[a] The cut-off value for the minimal number of matched vectors.
[b] The cut-off value for the minimal number of matched vectors within an asymmetric unit.
[c] The criterion for good solutions is that the angular distances of the solutions from the correct target structure are not greater than 15 degrees and the ranks of the solutions are not greater than 1000.
[d] The criterion for correct solutions is that the angle distances of the solutions from the correct target structure are not greater than 15 degrees.

4. Discussion

In our previous work,[11] we described a rotation function calculated by matching the search model directly with the Patterson vector map. The signal-to-noise ratio for the correct, meaningful match was increased by averaging all the matching peaks. Several matching scores were employed to evaluate the goodness of matching. These matching scores were then combined into a single total score by optimizing a set of weights from linear regression. It was shown that for a specific search model and a given crystal structure, there exists such a set of optimal weights that can be applied to the global rotation search and the correct solution can be ranked in the top 100 or less. However, this set of optimal weights in general cannot be transferred and applied to other search models and crystal structures with different space groups and cell parameters. For example, the set of weights optimized for 1a6m did not work for 1dti, that is, when the weights were applied to ranking global rotation solutions, the best rank for the correct solution was over 104. In this work, we try to solve this problem by designing a dynamic correlation coefficient. We also introduce two new matching scores: relative entropy and mean-square-residues. As shown in Table 2, the results are very encouraging. It shows that the dynamic correlation coefficient works for a variety of space groups and cell parameters in the global rotation search.

Relative entropy is also called Klieber–Kullback distance between two distributions and in special cases it is equivalent to maximum likelihood function.[20] So it may be crucial to include this measure of difference in evaluating the matching score. Mean-square-residue is another important measure of difference, which is sensitive to the difference between two distributions in different ways. However, the detailed

behaviour of these measures relative to search models and crystal forms has not been tested.

From Table 2, the best parameters for grid size, Nsum_cut-off_unit, and Nsum_cut-off_asymmunit seem to change with crystal forms slightly. For some combinations of parameters, there might be no good solutions as defined by our criteria (data not shown, see Table 2 for the definitions of 'good solution' and 'correct solution'). So it seems that these parameters need to be optimized. We consider that the dynamic correlation coefficient of the random solutions is the background noise and that of the correct solution the signal. Then the rank of the correct solutions is used as an indicator of the signal-to-noise ratio. It was found that the ranks of the correct solutions correlated weakly (cc∼0.32) with the averaged background noise obtained by sampling 30 random rotations. Therefore, in blind case, the background noise can be used as an indicator of the goodness of the parameters. That is, we optimize the adjustable parameters so that they give the lowest background noise obtained by sampling 30 random rotations globally. Listed in Table 3 are the number of good solutions and the best ranking of correct solutions obtained from global searches using the optimized parameters.

Furthermore, to have a clear view of how the new rotation function has been improved by using a dynamic correlation as well as two new scoring functions and optimized parameters, we made a comparison of the new rotation function and Phaser_BRF[2] (The search model is the same as in the previous jobs but has been rotated by a polar angle (45, 60, 180) relative to the target crystals). The results are listed in Table 4. It can be seen that our new rotation function has found correct solutions for all nine target crystals and six of them have at least one good solution, whereas only two crystals have correct solutions after Phaser_BRF searches. It suggests that the improvement of the new rotation function is more robust and more independent of the specific crystal structures.

Table 3. Number of good solutions and best ranking of correct solutions for global searches using the optimized parameters[a].

crystal	space group	asymmetric unit	grid size/Å	Nsum cut-off for matching	Nsum cut-off for asymmetric unit	number of good solutions	best ranking of correct solutions
1a6m	$P2_1$	1.0 1.0 0.5	5.2	2	8	1	937
1dti	$P2_12_12_1$	1.0 0.5 0.5	5.2	2	12	0	1899
102m	$P6$	0.5 0.5 1.0	5.4	4	8	1	6
1j3f	$P2_12_12_1$	1.0 0.5 0.5	4.8	4	8	1	841
1mlu	$P6$	0.5 0.5 1.0	5.0	2	8	1	457
1wvp	$P4_32_12$	0.5 0.5 1.0	4.8	2	12	2	670
2evp	$P2_12_12_1$	1.0 0.5 0.5	4.8	2	8	1	175
2jho	$P2_1$	1.0 1.0 0.5	4.8	2	12	5	39
2mm1	$P3_221$	1.0 1.0 0.5	5.2	4	12	1	227

[a] The selection of the adjustable parameters is based on the principle of the lowest average value of the dynamic correlation coefficient of 30 random rotations. The definitions of 'good solution' and 'correct solution' are given in Table 2.

Table 4. Number of good solutions and best ranking of correct solutions for global searches using the new rotation function[a] and Phaser_BRF.

crystal	number of good solutions for new rotation function	best ranking of correct solutions for new rotation function	number of good solutions for Phaser_BRF [b]	best ranking of correct solutions for Phaser_BRF
1a6m	1	24	0	*[c]
1dti	0	2743	0	*
102m	2	119	3	119
1j3f	1	695	0	*
1mlu	0	3822	3	21
1wvp	1	220	0	*
2evp	0	1650	0	*
2jho	2	136	0	*
2mm1	2	258	0	*

[a] The parameters have been optimized before the global searches of the new rotation function. The definitions of 'good solution' and 'correct solution' are given in Table 2.

[b] The global searches of Phaser_BRF are run with all default parameters except selecting to output all solutions.

[c] $*$ means no correct solution can be found in all of the output solutions.

References

[1] Murshudov G, von Delft F and Ballard C 2008 *Acta Cryst.* D **64** 0

[2] Storoni L C, McCoy A J and Read R J 2004 *Acta Cryst.* D **60** 432

[3] Keegan R M and Winn M D 2008 *Acta Cryst.* D **64** 119

[4] Trapani S and Navaza J 2008 *Acta Cryst.* D **64** 11

[5] Cohen S X, Ben Jelloul M, Long F, Vagin A, Knipscheer P, Lebbink J, Sixma T K, Lamzin V S, Murshudov G N and Perrakis A 2008 *Acta Cryst.* D **64** 49

[6] Schwarzenbacher R, Godzik A and Jaroszewski L 2008 *Acta Cryst.* D **64** 133

[7] Lebedev A A, Vagin A A and Murshudov G N 2008 *Acta Cryst.* D **64** 33

[8] Evans P and Mccoy A 2008 *Acta Cryst.* D **64** 1

[9] Kissinger C R, Gehlhaar D K and Fogel D B 1999 *Acta Cryst.* D **55** 484

[10] Glykos N M and Kokkinidis M 2000 *Acta Cryst.* D **56** 169

[11] Jiang F 2008 *Acta Cryst.* D **64** 561

[12] Vojtechovsky J, Chu K, Berendzen J, Sweet R M and Schlichting I 1999 *Biophys. J.* **77** 2153

[13] Liong E C, Dou Y, Scott E E, Olson J S and Phillips G N 2001 *J. Biol. Chem.* **276** 9093

[14] Ueno T, Koshiyama T, Ohashi M, Kondo K, Kono M, Suzuki A, Yamane T and Watanabe Y 2005 *J. Am. Chem. Soc.* **127** 6556

[15] Carlson M L, Regan R, Elber R, Li H Y, Phillips G N, Olson J S and Gibson Q H 1994 *Biochemistry* **33** 10597

[16] Kamiya N, Shiro Y, Iwata T, Iizuka T and Iwasaki H 1991 *J. Am. Chem. Soc.* **113** 1826

[17] Qin J, Perera R, Lovelace L L, Dawson J H and Lebioda L 2006 *Biochemistry* **45** 3170

[18] Arcovito A, Benfatto M, Cianci M, Hasnain S S, Nienhaus K, Nienhaus G U, Savino C, Strange R W, Vallone B and Della Longa S 2007 *P. Nat. Acad. Sci.* **104** 6211

[19] Hubbard S R, Hendrickson W A, Lambright D G and Boxer S G 1990 *J. Mol. Biol.* **213** 215

[20] Baldi P and Brunak S 2001 *Bioinformatics: the Machine Learning Approach* (Cambridge: The MIT Press) pp. 67–80