Acta Cryst. (1987). A**43**, 1-5

Generalized Method of Determining Heavy-Atom Positions Using the Difference Patterson Function

BY THOMAS C. TERWILLIGER* AND SUNG-HOU KIM

Department of Chemistry, University of California, Berkeley, California 94720, *USA*

AND DAVID EISENBERG

Department of Chemistry and Biochemistry and Molecular Biology Institute, University of California, Los Angeles, California 90024, *USA*

(*Received* 27 *August* 1984; *accepted* 16 *April* 1986)

Abstract

An automated procedure for locating the positions of heavy atoms in crystals of macromolecules has been developed. The method is simple to apply, is independent of space group, and permits inclusion of non-crystallographic symmetry. The procedure is a search of the difference Patterson function; a trial solution consisting of a set of heavy-atom sites is considered likely to be correct only if the corresponding 'minimum function' (the minimum value of the difference Patterson function at the self- and cross-vector positions for this group of sites) is large. Although the method may be used to search for 'single-site' solutions to the difference Patterson function, it is more effectively used to search for pairs of sites unrelated by crystallographic symmetry. In the latter case the number of predicted cross vectors for each trial solution is larger, and correct solutions may be more readily distinguished from incorrect ones. Because of noise in difference Patterson functions, it is helpful in evaluating the solutions obtained to calculate the probability that a given value of the minimum function might occur by chance. The method has been applied to nine difference Patterson functions for which solutions were known. In eight of these cases, including several which had resisted earlier attempts at interpretation, this procedure yielded at least half of the known sites.

Introduction

A crucial and often difficult step in the multiple isomorphous replacement method is the determination of the positions of heavy atoms in a crystal. The procedure most commonly used for this is inspection of a difference Patterson function (see Blundell & Johnson, 1976, for example). Unless there are only a few major heavy-atom sites in the asymmetric unit of a crystal, however, interpretation of a difference Patterson function may be difficult. This is due partly to the complexity of a Patterson function for many atomic sites. It is also due to a more subtle factor. *Difference* Patterson functions have an *intrinsic* level of noise that is roughly proportional to the number of heavy-atom sites. In contrast, the value of the difference Patterson function at points corresponding to vectors between heavy-atom sites is *independent* of the number of sites. As the number of heavy atoms increases, then, the noise level in a difference Patterson function increases, while the signal remains the same.

When interpreting difference Patterson functions for more than a few heavy-atom sites, the low ratio of signal to noise leads to two related problems. First, because there are many peaks in the difference Patterson function, many must be screened to find the one set corresponding to a correct solution. Second, discrimination of correct solutions from incorrect solutions based on spurious peaks may present substantial problems.

Single-site search procedures

The multiple-peak difficulty may be substantially reduced by using automated heavy-atom search techniques (Argos & Rossmann, 1976). The following single-site search is the basis of all our search methods.

The search is based on a 'minimum function', similar to that suggested by Buerger (1970). Given a *trial* set of atomic sites, all interatomic vectors may be calculated. If the trial solution is indeed a subset of the true structure, the difference Patterson function should usually have large values $\rho(\mathbf{u}_i)$ at all of the predicted interatomic vectors $\mathbf{u}_i$. We show below that the noise level $\sigma(\mathbf{u}_i)$ at a predicted vector $\mathbf{u}_i$ in a difference Patterson function can be estimated from the overall r.m.s. value of this function. A measure

* Present address: Department of Biochemistry and Molecular Biology, University of Chicago, 920E 58th Street, Chicago, Illinois 60637, USA.

0108-7673/87/010001-05$01.50

of the correspondence between a given trial set and the observed difference Patterson function is the smallest value of the ratio $R_i = \rho(\mathbf{u})/\sigma(\mathbf{u}_i)$ evaluated at the various predicted interatomic vectors for the set. A high minimum value indicates a good fit of the trial set to the difference Patterson function.

This approach, however, yields no measure of the likelihood that a given solution is based on spurious peaks in the difference Patterson function. Some additional information is required in order to do this.

Identification of trial solutions unlikely to be due to spurious peaks in the difference Patterson function

One method of identifying solutions likely to be correct is to assess, for each trial solution, the probability that a solution with its associated value of the minimum function (or higher) can occur by chance. If such an event is very unlikely then we may be confident that this solution is not based on 'noise' peaks in the difference Patterson function.

In the case of a single-atom search, the probability of finding, by chance, a trial solution with a given value R of the minimum function can be calculated in the following manner. Suppose the r.m.s. noise as a function of position in a difference Patterson function, $\sigma(\mathbf{u}_i)$, is known. First set up a grid of trial points in the asymmetric unit of the crystal; each of these trial points is to be tested as a potential single-site solution to the difference Patterson function. Then we may calculate the probability P that, purely by chance, at least one of these test points will have a corresponding value of the minimum function at least as large as, say, R_0. It is simplest to do this by first calculating the probability that *all* of the trial points will have a value of the minimum function *less* than R_0.

Consider a particular test point $\mathbf{x}$. This point in the real cell is associated with, say, M unique self vectors $(\mathbf{u}_i, i = 1, M)$, where we exclude the origin. The probability P_0 that the ratio $R_i = \rho(\mathbf{u}_i)/\sigma(\mathbf{u})$ at a particular self vector $\mathbf{u}$ will, by chance, be at least equal to R is given by the error function:

$$P_0 = 2\pi^{-1/2} \int_{R_0}^{\infty} \exp(-s^2/2)\,\mathrm{d}s. \qquad (1)$$

Then the probability that the ratio R_i will be at least equal to R_0 for all M Harker vectors is P_0 raised to the power M. That is, the probability that, by chance, a particular test point $\mathbf{x}$ will be associated with a value of the minimum function at least as large as R_0 is equal to P_0^M.

In this search procedure, many test points $\mathbf{x}_i$ will be considered as trial solutions. If N *well separated* positions are examined, the probability P that at least one of these is associated, by chance alone, with a value of the minimum function at least as large as R_0 can be calculated, assuming that all N tests are statistically independent. This yields

$$P = 1 - (1 - P_0^M)^N. \qquad (2)$$

If a finer grid is used so that N is very large, (2) is not strictly applicable. Consider two test points which are close together compared with the effective resolution of the difference Patterson function. The Harker vectors $\mathbf{u}_i$ associated with these two test points will be nearly identical, so that the values of the minimum function at these two points will be essentially equal. Therefore the probability P of finding a point $\mathbf{x}$ associated with a large value of the minimum function does not, as (2) would predict, increase to unity as the number of grid points in the search is increased indefinitely. In fact, once all the self vectors corresponding to adjacent points on the search grid are separated by much less than the effective resolution of the difference Patterson function, increasing the number of grid points can be expected to affect P only negligibly. We suggest that when a very fine search grid is used, an appropriate value of N to use in (2) is the maximum number of points in the asymmetric unit of the crystal such that no two of these points are associated with the same set of self vectors, within the resolution of the difference Patterson function.

The effective resolution of a difference Patterson function is not necessarily the limit of resolution of the data used to calculate it, because the intensities of reflections often decrease rapidly with increasing resolution. Perhaps a more realistic estimate of the resolution is the typical separation between maxima and minima in this function. Accordingly, we estimate the effective resolution of a difference Patterson function by dividing the volume of the asymmetric unit of this function by the number of local maxima and minima within the asymmetric unit. The length of an edge of a cube with this volume is taken to be the effective resolution.

The number of self vectors M associated with each trial point, the effective number N of trial points examined in a given search, and the noise level $\sigma(\mathbf{u})$ (see below) are in general simple to estimate and are essentially fixed for a particular difference Patterson function. Consequently, (1) and (2) may readily be used to evaluate potential solutions. A trial solution associated with a value R_0 of the minimum function is likely to be correct if the probability [P, given in (2)] of obtaining this value by chance is small (*e.g.* $P < 0{\cdot}05$).

Occasionally some of the self vectors $\mathbf{u}_i$ associated with a trial point may lie close to one another. In this case, it is necessary to have a criterion for determining whether the values of the difference Patterson function at these points are *necessarily* similar. We assume that two vectors are associated with independent values of the difference Patterson function if they are separated by more than one unit of the grid used to calculate the function.

Increasing the number of vectors predicted in the difference Patterson function for each trial point in a search

As shown by (2), the probability of finding an incorrect solution which has M cross vectors and a value of the minimum function greater than or equal to R_0 decreases very rapidly with increasing values of M. That is, the greater the number of peaks which must be present in the difference Patterson function before a trial solution is accepted, the less the likelihood of accepting an incorrect one. Consequently any factor that increases M (without greatly increasing N) will lead to an increased sensitivity in the search procedure.

I. *Non-crystallographic symmetry*

M increases if it is possible to include non-crystallographic symmetry in calculating 'equivalent' sites in the unit cell (Argos & Rossmann, 1976). If crystallographic and non-crystallographic symmetry are both included, M will increase from $N_{\text{equiv}}-1$ to $N_{\text{non}}N_{\text{equiv}}-1$, where N_{equiv} is the number of positions equivalent by space-group symmetry in the unit cell, excluding centering, and N_{non} is the number of positions equivalent by non-crystallographic symmetry alone in the unit cell, excluding any pseudo-centering. Non-crystallographic symmetry does not generally affect N.

II. *Searches for pairs of sites*

A second way to increase M, the number of distinct vectors in the difference Patterson function associated with each trial position in a search for heavy-atom sites, is to search for pairs of sites related by a fixed cross vector $\mathbf{y}$. Given $\mathbf{y}$ and the position of one trial site $\mathbf{x}_1$, the second site is located at $\mathbf{x}_2=\mathbf{x}_1+\mathbf{y}$. A search, similar to those described above, may be carried out. The value associated with each point $\mathbf{x}_1$ is the minimum function at (a) the self vectors associated with the points $\mathbf{x}_1$ and $\mathbf{x}_2=\mathbf{x}_1+\mathbf{y}$, and ($b$) all the cross vectors between positions equivalent to $\mathbf{x}_1$ and those equivalent to $\mathbf{x}_2$. Excluding the cross vector $\mathbf{y}$ which is common to all the points in the search, this generally yields $M=3N_{\text{equiv}}-3$ vectors predicted in the difference Patterson function for each test point.

In principle, it is necessary to carry out a search over both $\mathbf{x}_1$ and $\mathbf{y}$ in order to examine all possible pairs of sites. This very impractical search is not necessary, because a pair of sites $\mathbf{x}_1$ and $\mathbf{x}_2=\mathbf{x}_1+\mathbf{y}$ is unlikely to correspond to a correct solution unless the difference Patterson function is reasonably large at the cross vector $\mathbf{y}=\mathbf{x}_2-\mathbf{x}_1$. A very useful search procedure, then, is (a) to locate the isolated peaks in the difference Patterson function, some of which are likely to correspond to cross vectors between sites, and (b) to use the coordinates of each of these peaks, one at a time, as trial cross vectors $\mathbf{y}$ in a two-site search. Any solutions obtained may be evaluated, as before, on the basis of the probability of finding such a solution by chance.

The value of N in (2) for a two-site search may be estimated by analogy with the procedure used in a single-site search: N is the maximum number of points in the asymmetric unit of the crystal such that no two of these points are associated with the same set of self *and* cross vectors, within the effective resolution of the difference Patterson function.

Locating additional heavy-atom sites given a partial solution

Once a partial solution to a difference Patterson function has been obtained, a simple search may be carried out to locate additional sites, one at a time. At each trial position for a new site, the minimum function at the associated self vectors and at the cross vectors with the 'known' sites is recorded. The position with the largest value of this function is considered as a potential additional site. As before, the probability of finding an incorrect trial site with this value of the minimum function may be evaluated, and if it is sufficiently unlikely, the additional site may be added to the list of 'known' sites.

'Noise' in difference Patterson functions

Difference Patterson functions contain a considerable amount of what might be called 'noise', at least from the point of view of determining the locations of heavy-atom sites in the crystals. This noise is partly due to errors in measurement and scaling of native and derivative structure-factor amplitudes and to non-isomorphism between native and derivative structures. More important is that, for acentric reflections, the magnitude of the difference between native and derivative structure-factor amplitudes is not in general equal to the heavy-atom structure-factor amplitude (Blundell & Johnson, 1976). This introduces a high intrinsic noise level in all *difference* Patterson functions calculated from acentric diffraction data. The purpose of this section is to show that, in most cases, the r.m.s. value of this 'noise' in a region of a difference Patterson function is very similar to the r.m.s. value of the difference Patterson function itself in this region.

In the discussion below, we ignore the 'centric' reflections used to calculate difference Patterson functions, as there are generally few of them relative to the number of 'acentric' reflections.

In order to estimate the 'noise' level in a difference Patterson function, notice that if the magnitude of the true heavy-atom structure factor f_H for a reflection **h** is small relative to the observed native structure-factor amplitude $F_{p,\mathrm{obs}}$, if there is perfect isomorphism between native and derivative structures, and if there are no errors in measurement of native and derivative structure-factor amplitudes, we may write

$$F_{PH} - F_P \doteq f_H \cos\beta \qquad (3)$$

where β is the difference in phase angle between f_H and the native structure factor. Then, using (3), we obtain an expression for the coefficients of a difference Patterson function:

$$C_{\mathbf{h}} \equiv (F_{PH} - F_P)^2 \doteq f_H^2 \cos^2(\beta) \qquad (4a)$$

for each **h**, where the difference Patterson function $\rho(\mathbf{u})$ is given by

$$\rho(\mathbf{u}) = \sum_{\mathbf{h}} C_{\mathbf{h}} \cos(2\pi\mathbf{h}\,.\,\mathbf{u}). \qquad (4b)$$

The term with $\mathbf{h} = (0, 0, 0)$ is not included here or elsewhere in this work. Notice that the true Patterson function for the heavy-atom sites is based on coefficients $C_{\mathbf{h}} = f_H^2$. Rearranging (4*a*) we obtain

$$C_{\mathbf{h}} \simeq f_H^2/2 + [\cos^2\beta - 1/2]f_H^2. \qquad (5)$$

The first term on the right-hand side of (5) corresponds to the 'signal' in the difference Patterson function; it is half the 'ideal' value of the signal (f_H^2). The second term, uncorrelated with the first, since $[\cos^2\beta - 1/2]$ has a mean value of zero and varies in a 'random' fashion, corresponds to the 'noise'. Using (4*a*) and (4*b*), we can deduce that the r.m.s. value of a difference Patterson function with M terms is

$$\rho_{\mathrm{r.m.s.}} \sim M^{1/2}(\langle f_H^4\rangle^{1/2})(3^{1/2}/4) \qquad (6a)$$

where the angle brackets indicate an average value and we have used the fact that $\langle\cos^4\beta\rangle = 3/8$ for random angles β. The r.m.s. values of the signal (S) and the 'noise' (σ) in this function may be estimated in a similar fashion using the first and second terms on the right-hand side of (5) respectively, yielding:

$$S_{\mathrm{r.m.s.}} \sim M^{1/2}(\langle f_H^4\rangle^{1/2})(2^{1/2}/4) \qquad (6b)$$

and

$$\sigma_{\mathrm{r.m.s.}} \sim M^{1/2}(\langle f_H^4\rangle^{1/2})(1/4). \qquad (6c)$$

Therefore the ratio of signal to noise is $S_{\mathrm{r.m.s.}}/\sigma_{\mathrm{r.m.s.}} \simeq 2^{1/2}$ and the ratio of the overall r.m.s. value of the difference Patterson function to the r.m.s. value of the noise is $\rho_{\mathrm{r.m.s.}}/\sigma_{\mathrm{r.m.s.}} \simeq 3^{1/2}$. The intrinsic 'noise' is entirely due to the fact that the phase-angle difference β is not always 0 or π. In the case where substantial errors in measurement of native and derivative structure factors exist, or where some lack of isomorphism between native and derivative structures is present, $\sigma_{\mathrm{r.m.s.}}$ is even closer to $\rho_{\mathrm{r.m.s.}}$. It therefore is never very incorrect to take the overall r.m.s. value of a difference Patterson function as an estimate of the overall r.m.s. error in this function. Notice that $\langle f_H^2\rangle$ and therefore $\sigma_{\mathrm{r.m.s.}}$ (6*c*) is roughly proportional to the number of heavy-atom sites.

The 'noise' level is not the same everywhere in the difference Patterson function, however. An analysis similar to that used to obtain (6*c*) may be used to show that this equation applies to zones within the asymmetric unit of the difference Patterson function as well as to overall values. The 'noise' level does, however, depend on the point symmetry of the position in the difference Patterson function. If the number of elements in the point symmetry at a particular position is L, it may be shown that the r.m.s. 'noise' will be roughly $L^{1/2}$ times that at general positions. For example, at a position of mirror symmetry in a difference Patterson function, the r.m.s. noise is 1·4 times that at general positions.

Discussion

The procedure we have followed when applying the methods described here is: (*a*) to carry out a search for single-site solutions to the difference Patterson function, noting solutions which are very unlikely to occur by chance (*i.e.* with a probability of less than 0·05) and (*b*) to carry out searches for two-site solutions using each of one to 30 isolated peaks in the difference Patterson function as the fixed cross vector in these searches. If any two-site solutions found are very unlikely to have occurred by chance, a search for additional sites is carried out with these sites serving as starting point. These additional sites which are very unlikely to be due to chance are included in the reported solútion. If more than one solution was obtained using this procedure, only the solution which was least likely to be due to chance was considered. No more than six sites were considered at a time.

Table 1 summarizes the results of this procedure as applied to nine heavy-atom derivatives of four structures for which heavy-atom sites were known from earlier work. In eight of the nine cases, our procedure clearly identified at least half of the major sites in the known solution. These correctly identified sites were a r.m.s. distance of 1·5 Å from the corresponding known sites. In some cases, the solutions obtained contained additional sites which were consistent with the difference Patterson function. Three of the difference Patterson functions which were correctly solved by the present methods had resisted earlier attempts at interpretation by inspection.

Our procedure failed in one case, the samarium (Sm in Table 1, Kim *et al.*, 1972) derivative of *t*RNA$_{\mathrm{phe}}$. Two solutions, each consisting of six atomic sites, were found which were very unlikely ($P <$ 0·001) to have occurred by chance, yet which had no sites in common with the two reported sites (Kim *et al.*, 1972) obtained by difference Fourier analysis. In

Table 1. *Application of search procedures to difference Patterson functions with known solutions*

Crystal structure	Melittin		Monellin		Yeast *t*RNA$_{\text{phe}}$			Ribulose bisphosphate carboxylase	
Derivative	Hg	KI	Au	Pd	Os	Pt	Sm	Pt	DMM
Previously solved using difference Patterson?	Yes	No	Yes	No	Yes	No	No	Yes	Not tried
Number of known sites	1	5	2	4	1	3	2	1	4
Number of correct sites found (present method)	1	5	2	2	1	2	0	1	3
R.m.s. distances from known sites (Å)	0·3	0·9	2·2	1·1	0·4	2·5	—	2·9	0·7
Number of additional sites found	0	1	3	0	0	1	6	1	1
Computation time (min):									
FFT:	2	2	1	1	5	5	5	10	10
search:	4	4	15	15	10	54	37	10	10
Space group	*C*222$_1$		*P*2$_1$		*P*2$_1$22$_1$			*I*422	
Unit cell dimensions (Å)	61 × 38 × 42		40 × 72 × 87		36 × 56 × 162			149 × 149 × 138	
Resolution (Å)	2·8		5·0		5·0			4·0	
Reference	Terwilliger & Eisenberg (1982)		Tomlinson & Kim (1981)		Kim, *et al.* (1972)			Baker, Suh & Eisenberg (1977)	

fact, our procedure could not have obtained the correct solution in this case, as each of the two sites in the known solution had at least one associated self vector in a position on the difference Patterson function which has a value far below zero. We do not know if either solution obtained by our program consists of correct minor sites or if they are completely incorrect.

An important feature of the two-site search procedure is that it yields sets of sites which, as a group, form a solution to the difference Patterson function. In contrast, a single-site search only yields a list of potential solutions, determined only to within an 'origin shift' which depends on the symmetry of the structure. Crystal structures with the symmetry of space group *P*222, for example, have in general eight non-equivalent positions in the unit cell, which share a given set of associated self vectors. The two-site search procedure yields sets of sites which are all related to the same origin and are all part of the same enantiomer of the true solution.

Although non-crystallographic symmetry operations were not applied in any of the cases listed in Table 1, when local symmetry is present and has been characterized it may be readily applied in using any of the methods described here.

The procedures we have discussed here could also be applied to Patterson functions (as opposed to difference Patterson functions) which are based on structures containing a small number of atoms of equal size or a small number of heavy atoms along with many smaller atoms. In the case of Patterson functions which contain little 'noise', however, it should be noted that the methods described here for identifying the solutions unlikely to be due to chance lose their usefulness, as nearly all features of the difference Patterson function will correspond to some element of the true structure in question.

Only a small amount of computation time is required to use the methods described here; in the cases we have examined so far, the time required to obtain the solutions listed in Table 1 was from one to fourteen times that required to calculated the asymmetric unit of the difference Patterson function using a fast-Fourier-transform method and a grid corresponding to one-third to one-sixth the resolution of the reflection data. We suggest that, although one should always visually examine a difference Patterson function in order to detect unusual features such as local symmetries and unusually shaped peaks, the automated procedures described here may conveniently be used as first attempts to determine solutions to complicated or simple difference Patterson functions, rather than as a last resort.

A Fortran program (*HASSP*) which incorporates the results described here may be obtained from the Protein Data Bank, Brookhaven National Laboratory, Upton, Long Island, New York, 11973, USA.

We thank W. Smith and C. Ogata for the use of their data and S. Holbrook for useful comments. Support of USPHS NIH GM 31299 and GM 29287 are gratefully acknowledged.

References

ARGOS, P. & ROSSMANN, M. G. (1976). *Acta Cryst.* **B22**, 2975–2979.

BAKER, T. S., SUH, S. W. & EISENBERG, D. (1977). *Proc. Natl Acad. Sci. USA*, **74**, 1037–1041.

BLUNDELL, T. L. & JOHNSON, L. N. (1976). *Protein Crystallography.* New York: Academic Press.

BUERGER, M. J. (1970). *Contemporary Crystallography.* New York: McGraw-Hill.

KIM, S. H., QUIGLEY, G., SUDDATH, F. C., MCPHERSON, A., SNEDEN, D., KIM, J. J., WEINZIERL, J., BLATTMANN, P. & RICH, A. (1972). *Proc. Natl Acad. Sci. USA*, **69**, 3746–3750.

TERWILLIGER, T. C. & EISENBERG, D. (1982). *J. Biol. Chem.* **257**, 6010–6015.

TOMLINSON, G. E. & KIM, S. H. (1981). *J. Biol. Chem.* **256**, 12476–12477.