Proc. Indian Acad. Sci. (Chem. Sci.), Vol. 92, Numbers 4 & 5, August & October 1983, pp. 329–334.
© Printed in India.

An assessment of criteria of fit in Patterson search

C E NORDMAN

Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109, USA

Abstract. The problem of reliably detecting a known set of n vectors of weight w_i embedded in a heavily overlapped Patterson function P_i is investigated by a Monte-Carlo simulation based on searches of computer-generated random number sequences. Several formulations of the criterion of fit were compared. All were found to improve when the criterion was based on a subset of m "worst fitting" vectors as judged by a low value of (P_i/w_i). The best criteria were $\Sigma_{i=1}^{m}(w_iP_i)/\Sigma_{i=1}^{m}w_i^2$, with $m \approx (0{\cdot}4–0{\cdot}5)n$, $\Sigma_{i=1}^{m}P_i/\Sigma_{i=1}^{m}w_i$,with $m \approx 0{\cdot}3n$, and $\Sigma_{i=1}^{m}(w_iP_i)$ with $m \approx 0{\cdot}7n$. In each case the detectability of the embedded vectors w_i increases with increasing $\sigma(w)$ in relation to $\sigma(N)$, the standard deviation of the overlaid noise. A related simulation of a Patterson search for non-crystallographic symmetry shows that for a given size of the non-crystallographically symmetric region, the detectability increases with the order (2-fold, 6-fold, 12-fold) of the symmetry.

Keywords. Patterson search; signal detection; Monte-Carlo simulation.

1. Introduction

Crystallographic structure-solving techniques based on interpretation of the Patterson function typically involve computer-implemented systematic sampling or exhaustive search of the Patterson function. The manner in which the Patterson function is sampled reflects the nature of the available *a priori* information concerning the unknown crystal structure.

This information may be the knowledge of the structure of some (rigid) fragment present in the molecule. In this case the Patterson function is sampled at points corresponding to the set of vectors within this fragment, or between two properly oriented copies of the fragment. Alternatively, the available information may be the knowledge that the molecule—perhaps a multisubunit protein, or a virus—possesses local non-crystallographic symmetry of a particular kind. In this case the Patterson function may be sampled with a movable "symmetry grid," an array of points in spherical polar coordinates, which possesses the exact point group symmetry of the molecules (Nordman 1980a).

A recent review (Nordman 1980b) describes these methods in greater detail and gives examples of their use in small-molecule crystallography.

In either type of search a criterion of fit, or "image-seeking function," is evaluated at each step in the search. Collectively, these values constitute a "map," generally in three dimensions, the coordinates being angular or translational depending on the nature of the search. Promising orientations or translational positions of the search object are indicated by maxima, or minima, in the map.

In small-molecule problems, where the rigid fragment constitutes most or all of the molecule, the exact formulation of the criterion of fit is not very crucial. Criteria used in such cases include maximizing the sum of the sampled values of the Patterson function (Braun *et al* 1969), or minimizing the sum of the squares of the difference between the fragment Patterson and the crystal Patterson at all points where the former, unacceptably, exceeds the latter (Huber 1965).

The problem of making the search as discriminating as possible was first considered by Schilling (1970). He showed that it is advantageous to sort the sampled Patterson values P_i and the weights w_i of the known sampling vectors, in order of increasing values of the ratio P_i/w_i. The lower the value of this ratio, the worse is the fit, that is, the more poorly is the fragment vector peak w_i accommodated by the Patterson value P_i. By including in the criterion of fit only those (w_i, P_i) values which are most discriminating, as indicated by low P_i/w_i, Schilling showed that more reliable search results were obtained. The "minimum average" is defined as

$$\text{MIN}(m, n) = \sum_{i=1}^{m} P_i \Big/ \sum_{i=1}^{m} w_i \quad m < n, \tag{1}$$

where n is the total number of fragment vectors used, and m is the size of the subset having the m lowest values of P_i/w_i. This criterion of fit has found wide use, typically with $m/n = 0{\cdot}1$–$0{\cdot}3$.

Another reasonable criterion of fit is the sum of the products of the search vector weights w_i and the sampled Patterson values P_i. This function

$$\text{SP} = \sum_{i=1}^{n} w_i P_i \tag{2}$$

tends to be high at the correct solution. Recognizing that w_i represents points in a "model" Patterson, P_m, and replacing the sum with an integral, it is seen that (2) is related to $\int P_m(\mathbf{r})P(\mathbf{r})\,d\mathbf{r}$. This integral, evaluated as a sum of products of Patterson *coefficients*, is the criterion of fit used in reciprocal-space search methods, for example, the widely used rotation function (Rossmann and Blow 1962).

In order to assess the potential value of Patterson-space search techniques in macromolecular crystallography, it is of interest to examine different criteria of fit in the hope of finding the one which is most promising in the unfavourable case of a very heavily overlapped Patterson function. It has recently been shown (Nordman and Hsu 1982) that a Monte-Carlo calculation which simulates a Patterson search can be formulated as a one-dimensional problem of detecting a sequence of known numbers embedded in a longer sequence of random numbers. On the basis of relatively limited statistics it was concluded that the criterion (1), with $n = 300$, $m = 50$–100, was more successful in finding the "correct" solution than criterion (2).

In this communication a slightly modified Monte-Carlo calculation is used to examine several criteria of fit in the light of much more extensive statistical material. Also, a related formulation is employed in a Monte Carlo simulation of a search for local non-crystallographic symmetry.

2. Simulated structure search

In an actual Patterson search n vectors of weight w_i scan the Patterson function, returning, at each point in the search a set of values P_i, $i = 1, \ldots, n$. The arrays P_i and w_i are sorted in ascending order of (P_i/w_i) and criteria of fit based on the first m entries in the sorted arrays, where $m \leqslant n$.

A rapidly computable simulation of this is as follows. Let the "Patterson" to be searched be represented by $P_i = S_i + N_i$ where S_i and N_i are sequences of random numbers with positive means $\langle S \rangle$ and $\langle N \rangle$ and standard deviations $\sigma(S)$ and $\sigma(N)$. These sequences are of equal length l, here taken as 500.

The n search vectors w_j in this simulation are a positionally significant subset of the sequence S_i. Without loss of generality the w_j can be taken as a contiguous sequence $w_j = S_{j+k_0}$ where the $(k_0 + 1)$th entry in the S_i sequence is the first in w_j. The w_j's are treated as the "known" search vector weights; n was taken as 300 here. The search is carried out by translating the w_j sequence along the P_i sequence, allowing the w_j's to sample P_{j+k} for successive values of k, from zero to $l - n$. At each of the 201 steps in the search the data are sorted, and several criteria of fit evaluated. A given criterion is judged successful if it assumes a higher value when $k = k_0$ than for any of the 200 other values of k. The percentage of successes scored by a given criterion in a large number of independent searches allows us to compare different criteria with one another.

Five different criteria of fit were evaluated. These included the minimum average, MIN, as defined in eq. (1), a weighted minimum average

$$\textsc{wmin}(m, n) = \sum_{i=1}^{m} (w_i P_i) \Big/ \sum_{i=i}^{m} w_i^{\,2} \quad m \leqslant n \tag{3}$$

and the quantity

$$\textsc{msp}(m, n) = \sum_{i=1}^{m} (w_i P_i) \quad m \leqslant n \tag{4}$$

which is a generalization of SP (equation (2)). The criteria $\Sigma_{i=1}^{m} (1 + w_i/\langle w \rangle)P_i/\Sigma_{i=1}^{m} w_i$ and $\Sigma_{i=1}^{m} (P_i/w_i)$, $m \leqslant n$, were also computed. They were less successful than the others, and are not further discussed.

Figure 1 shows the results for the three best criteria of fit. In these calculations the noise, N_i, was taken as Gaussian with a mean of 300 and a standard deviation $\sigma(N) = 100$. The sequence S_i, including the vector weights w, was also Gaussian with a mean $= 3\sigma(S)$, and $\sigma(S)$ ranging from 15 to 50. For each choice of $\sigma(S)$, at least 100 runs were calculated.

The results are summarized in figure 1. It should be noted that the 'noise level' of the ordinate is 1/201 or 0·5 %; this would be the statistical chance of a 'successful' search with a vanishingly small $\sigma(S)$. At the upper end, $\sigma(S) \approx 0{\cdot}5\sigma(N)$ essentially insures success.

All curves reach maxima at values of $m < n = 300$. The position of the maximum appears to be independent of $\sigma(S)/\sigma(N)$ for any one criterion of fit.

With $m = 300$, the MIN criterion (equation (1)) reduces to $\Sigma_{i=1}^{300} P_i$ divided by a constant. Since the calculation is designed so that $\langle w \rangle \approx \langle S \rangle$, the MIN criterion is meaningless at $m = 300$, and the corresponding points are not shown.

The optimal choice of m for MIN can be estimated to be approximately 0·3 n, in agreement with past experience. The WMIN criterion tends to have its maximum at (0·4–0·5)n, and in essentially every run tends to give slightly higher success rates than MIN. The maxima of MSP lie at approximately 0·7 n, and tend to be lower than those of either WMIN or MIN. It should be noted that at $m = 300$ MSP is identical to SP (equation (2)). This criterion is distinctly inferior to any of the others.

Additional runs were done with N_i unchanged, but S_i *uniformly* distributed between 0·01 and variable upper limits, up to 60. When compared to the Gaussian runs with equivalent $\sigma(S)$ no clear difference in success rates could be discerned.

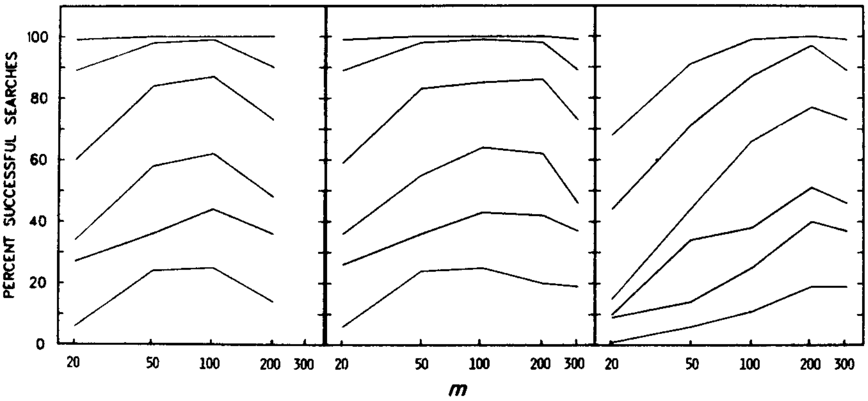

Figure 1. Percentage of successful structure searches as a function of $m \leqslant n = 300$ for three criteria of fit as defined in the text: MIN (left), WMIN (centre) and MSP (right). For each criterion the percentage is shown for $\sigma(S)/\sigma(N) = 0{\cdot}15$ (bottom graph), 0·2, 0·25, 0·3, 0·4 and 0·5 (top graph).

Finally the effect of varying the mean of the noise N_i was explored, keeping $\sigma(N)$ constant at 100. Lowering $\langle N \rangle$ to 250 caused some deterioration; raising it to 1000 or above also appeared to cause some decline. One may tentatively conclude that the optimal choice of the constant term in the Fourier synthesis of the Patterson is 3 to 5 times $\sigma(P)$; this choice also achieves a reasonable compromise between numerical accuracy and packing density in the computer memory.

3. Simulated symmetry search

In each simulated symmetry search a set of 50 independent random number sequences

$$F_i = S_i + N_i \quad i = 1, \ldots, 960$$

was searched. The noise sequences N_i were taken as Gaussian with $\langle N \rangle = 100$ and $\sigma(N) = 20$. The sequences S_i were also Gaussian with $\langle S \rangle = 0$ and $\sigma(S)$ variably chosen from 0·05 $\sigma(N)$ to 0·45 $\sigma(N)$.

One of the 50 S_i sequences was modified, so as to give it n_s-fold symmetry, by requiring that $S_{i+j} \equiv S_i$, where $j = 960/n_s$. Thus, this one S_i sequence consists of n_s copies of a random number sequence of length j.

For each of the 50 F_i-sequences the quantity

$$\sum_{i=1}^{j} \sum_{s=0}^{n_s-1} [F_{i+js} - \langle F_{i+js} \rangle_s]^2$$

was evaluated. This quantity is expected to have its lowest value for the "symmetric" sequence. If this is indeed found, the search is taken as successful. The "noise level" for this search, as formulated here, is 1/50 or 2%.

Symmetry searches were carried out for $n_s = 12$, 6 and 2. For each choice of $\sigma(S)$ 50 to 100 searches were done. The results, shown in figure 2 bear out the expected sharp increase in the success rate with increasing strength of the symmetric component, $\sigma(S)$,

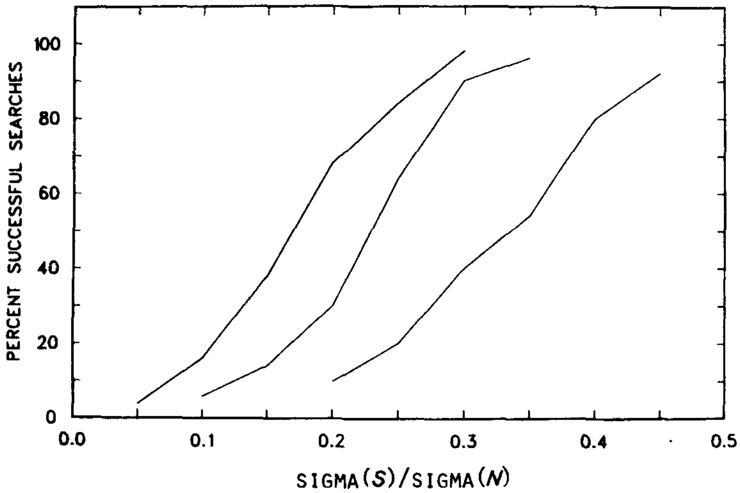

Figure 2. Percentage of successful symmetry searches as a function of $\sigma(S)/\sigma(N)$ for 12-fold (left curve), 6-fold (centre) and 2-fold (right) local non-crystallographic symmetry. The size of the noncrystallographically symmetric region is the same in the three cases.

in relation to the noise, $\sigma(N)$. The three curves also show the increase in detectability which accompanies an increase in symmetry, here from 2-fold to 6-fold and 12-fold.

4. Conclusions

The results unambiguously demonstrate that the discriminating power of a Patterson-space structure search improves when the "minimum average" principle is applied in calculating the criterion of fit. We conclude that in searches where the detectability of the correct solution is at all in doubt, the benefits of calculating minimum averages are well worth the additional computing time required for the sort, on P_i/w_i, which is carried out at every step in the search.

Somewhat more tentatively, the results suggest that the criterion

$$\sum_{i=1}^{m} w_i P_i \Big/ \sum_{i=1}^{m} w_i^2,$$

with $m \approx (0{\cdot}4 - 0{\cdot}5)n$, is superior to the presently used

$$\sum_{i=1}^{m} P_i \Big/ \sum_{i=1}^{m} w_i.$$

Both are consistently superior to the criterion

$$\sum_{i=1}^{n} (w_i P_i).$$

It should be emphasized that it has not been shown or suggested that any of these criteria is the best one that can be formulated. What the best one is, is still an open question.

References

Braun P B, Hornstra J, and Leenhouts J I 1969 *Philips Res. Rep.* **24** 85

Huber R 1965 *Acta Crystallogr.* **19** 353

Nordman C E 1980*a Acta Crystallogr.* **A36** 747

Nordman C E 1980*b Computing in crystallography* (eds) R Diamond, S Ramaseshan and K Venkatesan (Bangalore: Indian Academy of Sciences) p. 501

Nordman C E and Hsu L-Y R 1982 *Computational crystallography*, (ed.) D Sayre (New York: Oxford) p. 141

Rossmann M G and Blow D M 1962 *Acta Crystallogr.* **15** 24

Schilling J W 1970 *Crystallographic computing* (ed.) F R Ahmed (Copenhagen: Munksgaard) p. 115