The application of CLEAN to crystallographic problems

RAMESH NARAYAN

Raman Research Institute, Bangalore 560 080, India

Abstract. It is suggested that the radio astronomical technique of image reconstruction called CLEAN could be profitably used in crystallography to improve sharpened Patterson maps and E-maps. The method is here generalized to handle the symmetries which arise in crystallographic maps. Simulations on model one and two dimensional structure are presented to illustrate the power of the method.

1. Introduction

In radio astronomy, the two-dimensional sky brightness distribution is often measured by interferometric techniques. These experiments measure the Fourier transform of the required function, but only over a finite region of reciprocal space. Moreover, the experimental data are often not even simply connected but have 'holes' and 'sectors' of missing information in reciprocal space. The 2-D radio maps constructed from such data are difficult to interpret, both due to the lack of resolution as well as due to the termination ripple which can swamp many of the weaker features. In recent years, a method called CLEAN (Högbom 1974) has been widely used to improve the interpretability of radio maps. CLEAN is remarkably successful in removing ripples and side-lobes. It can also be used to increase the resolution within reasonable limits. One of the reasons for the success of CLEAN is that radio maps are as a rule 'essentially empty' *i.e.* the real space function one seeks (the true map) is zero or nearly zero over large areas and has significant non-zero values in small isolated regions. This appears to be a requirement for the successful implementation of CLEAN.

In this paper, it is suggested that there are many situations in crystallography where CLEAN can be profitably used. We may mention in particular the following two cases (see for instance Stout and Jensen 1968):

    (a) Sharpened Patterson maps, and

    (b) E-maps computed from phases obtained through the direct methods.

In both the above cases, the peaks in the true map are sharp and well separated with essentially zero density in between. However, in practice, termination ripple is a problem. Consequently, fully sharpened Patterson maps are often quite useless and E-maps sometimes have false peaks. We expect CLEAN to reduce ripple significantly in both maps, thus increasing their usefulness. It may even be possible to enhance the resolution (by as much as a factor of two).

We begin this paper with a brief description in § 2 of the CLEAN algorithm. We then describe in § 3 the elegant viewpoint of Schwarz (1978) who has interpreted CLEAN as a form of least squares fitting. Section 4 discusses the problem of adapting CLEAN to the symmetries which arise in crystallography (there is no symmetry in radio astronomy). We show that the approach of Schwarz can be used to suitably generalize CLEAN. Finally, we present one and two dimensional simulations in §5 which serve to bring out the potential of CLEAN in crystallographic applications.

2. The CLEAN algorithm

The following is a brief description of the CLEAN algorithm (Högbom 1974):

(a) Compute the 'dirty map' $d(r)$ using the incomplete set of Fourier coefficients $F(h_i)$. $d(r)$ is called the 'image' in optical terminology. Although $d(r)$ is a continuous function of the variable r, in practice it will generally be computed on a grid of points. Hence, we will deal with the sampled dirty map $d(r_i)$ with the provision that the sampling is sufficiently fine to handle any resolution which we might aim at.

(b) Compute $p(r_i)$, the 'point spread function' (PSF) (this is called the 'dirty beam' in radio astronomy). $p(r_i)$ is obtained by setting $F(h_i)$ to the value 1 at all measured points in the reciprocal space and zero elsewhere. We note that $p(r_i)$ is just the dirty map corresponding to the case when the true map consists of a single delta function of unit weight at the origin.

(c) Locate the grid point r_k at which the current dirty map attains its maximum value, say d_{max}. Postulate in the true map a δ-function at r_k of strength $gd_{max}/p\,(0)$ (where g is a number, called the gain factor, which is to be set by the user and is usually given a value in the range 0 to 1). Subtract from $d(r_i)$ a copy of the PSF of strength $gd_{max}/\,p\,(0)$ centred at r_k, to obtain the new dirty map

$$i.e.\ [d(r_i)]_{new} = [d(r_i)]_{old} - \frac{gd_{max}}{p(0)}p\,(r_i - r_k). \tag{1}$$

Simultaneously, add a δ-function of strength $gd_{max}/p\,(0)$ located at r_k to the 'components map' $t\,(r_i)$, which accumulates the δ-functions corresponding to the subtracted peaks.

(d) If the new dirty map still retains sufficient structure (which can be checked by testing if $d(r_i)$ is anywhere greater than some specified cut-off value), go to (c). If not, go to (e).

(e) Compute a 'clean peak' $h(r_i)$ (called the 'clean beam' in radio astronomy) which is centred at $r = 0$, has unit weight, has no or negligible ripple and has a half-width corresponding to the resolution required in the final map. A convenient choice for $h(r_i)$ is a Gaussian.

(f) Convolve the components map $t(r_i)$ with the clean peak $h(r_i)$ to get the 'clean map' $c(r_i)$. An option (generally exercised in radio astronomy) is to add the residual dirty map to $c(r_i)$.

The following general features of CLEAN should be kept in mind to appreciate and understand the method.

(a) CLEAN models the true map as a sum of δ-functions. Since the problem of solving for the δ-functions does not have a unique answer, a sequential scheme is

used. The method is obviously best suited to those problems where the true map is nearly a sum of δ-functions. Hence the expectation of success with sharpened Patterson maps and E-maps.

(b) Step (c) above, where the maximum in the dirty map is located, is intuitively appealing but may appear to be arbitrary. Schwarz (1978) has shown that this step can be given an elegant least squares interpretation, which we discuss in § 2.

(c) Because of the choice of the clean peak $h(\underline{r}_l)$, the clean map has no termination ripple. There is moreover the option to choose $h(\underline{r}_l)$ so as to increase the resolution of the map (within reasonable limits). This is likely to be useful in crystallographic applications.

(d) CLEAN extrapolates the Fourier data. This follows from the fact that the Fourier transform of the δ-components map $t(\underline{r}_l)$ extends upto infinity without decaying. The role of the clean peak is to reduce the extent of Fourier extrapolation to reasonable limits without introducing ripple.

3. Least squares interpretation of CLEAN

Schwarz (1978) has shown that CLEAN can be rationalized as a sequential least squares scheme. The argument described here (which is slightly different from that used by Schwarz) will be generalized in the next section for crystallographic applications.

It was mentioned in § 2 that CLEAN iteratively accumulates δ-functions in the components map by locating at each iteration the maximum in the dirty map. Let us investigate the following 'alternative' least squares strategy for selecting the δ-functions: Select the strength S and position $\underline{r}_k$ of the δ-function such that the residual dirty map after subtracting the corresponding PSF is *minimized in a least squares sense.* In other words, optimize S and $\underline{r}_k$ such that the following quantity R' is a minimum

$$R' = \sum_i [d(\underline{r}_l) - Sp(\underline{r}_l - \underline{r}_k)]^2. \tag{2}$$

We will show that the solution to (2) corresponds to the CLEAN choice discussed in step (c) in § 2.

Let $F(\underline{h}_l)$ be the Fourier transform of $d(\underline{r}_l)$, i.e.

$$d(\underline{r}_l) = \frac{1}{n} {\sum_j}' F(\underline{h}_j) \exp(-2\pi i \underline{h}_j \cdot \underline{r}_l), \tag{3}$$

where n is the total number of grid points and the prime indicates that the sum includes only those reciprocal points where measurements are available. The PSF $p(\underline{r}_l - \underline{r}_k)$ is clearly expressible as

$$p(\underline{r}_l - \underline{r}_k) = \frac{1}{n} {\sum_j}' \exp[-2\pi i \underline{h}_j \cdot (\underline{r}_l - \underline{r}_k)]$$

$$= \frac{1}{n} {\sum_j}' \exp(2\pi i \underline{h}_j \cdot \underline{r}_k) \exp(-2\pi i \underline{h}_j \cdot \underline{r}_l). \tag{4}$$

P.—3

Thus, the Fourier transform of $p(r_i - r_k)$ is $\exp(2\pi i\,h_j \cdot r_k)$. Now, by Parseval's theorem, minimizing R' in (2) is equivalent to minimizing R below.

$$R = \sum_j{}' |F(\underline{h}_j) - S\exp(2\pi i\,h_j \cdot \underline{r}_k)|^2, \tag{5}$$

where once again S and r_k are the parameters at our disposal. For a given $\underline{r}_k$, the optimization of R with respect to S gives

$$S = \frac{\sum_j{}' F(h_j)\exp(-2\pi i\,h_j \cdot r_k) + \sum_j{}' F^*(h_j)\exp(2\pi i\,\underline{h}_j \cdot \underline{r}_k)}{2\sum{}'\,l}$$

$$= d(\underline{r}_k)/p(0), \tag{6}$$

where we have used the well-known relation

$$F^*(\underline{h}_j) = F(-\underline{h}_j). \tag{7}$$

Substituting the expression for S given by (6) back in (5), we obtain the minimum value of R that can be attained for a fixed r_k to be

$$R_{\min}(\underline{r}_k) = \sum_j{}' |F(\underline{h}_j)|^2 - \frac{n}{p(0)}[d(r_k)]^2. \tag{8}$$

The optimization of $R_{\min}(\underline{r}_k)$ with respect to r_k is now trivial. Equation (8) obviously leads to a choice of that grid point r_k where the dirty map achieves its maximum value. Thus, the least squares scheme which we have just studied leads to the CLEAN prescription given earlier, except for the absence of the gain factor g, which is here equal to 1. The use of smaller values of g is a practical detail to avoid the danger of oversubtracting peaks (which is a problem whenever there is significant interference from neighbouring peaks).

4. CLEAN in the presence of symmetries

Crystallographic maps usually have symmetries (for instance, the Patterson is always centrosymmetric). Consequently, in crystallographic applications, it is not correct to sequentially remove *single* δ-functions. Peaks come in symmetry sets and one should therefore sequentially remove *sets* of symmetry related δ-functions. We therefore suitably generalize the CLEAN algorithm. To do this, we employ the least squares approach discussed in § 3.

For simplicity, let us assume that the only symmetry in the map is one of inversion at the origin. Then, for every peak at r_i, there is an identical one at $-r_i$. Consequently, in each iteration of CLEAN, we should select a pair of inversion-related δ-functions. The quantity R to be optimized now (corresponding to the expression (5) in the earlier case) is

$$R = \sum_j{}' |F(\underline{h}_j) - S[\exp(2\pi i\,\underline{h}_j \cdot \underline{r}_k) + \exp(-2\pi i\,h_j \cdot \underline{r}_k)]|^2. \tag{9}$$

As before, for a given r_k, we obtain the optimum value of S to be

$$S = \frac{2\sum_j' F(h_j) \exp(-2\pi i\, h_j \cdot r_k) + 2\sum_j' F^*(h_j) \exp(2\pi i\, h_j \cdot r_k)}{2\sum_j' [2 + \exp(2\pi i\, h_j \cdot 2r_k) + \exp(-2\pi i\, h_j \cdot 2r_k)]}$$

$$= \frac{d(r_k)}{p(0) + p(2r_k)}, \tag{10}$$

where we have used the symmetry of the dirty map and the PSF

i.e., $d(r_k) = d(-r_k); \quad p(2r_k) = p(-2r_k).$ (11)

Equation (8) is now generalized to

$$R_{\min}(r_k) = \sum_j' |F(h_j)|^2 - n\frac{[d(r_k)]^2}{p(0) + p(2r_k)}. \tag{12}$$

Now we see that it is not the maximum in the dirty map itself which is important, but rather the maximum in the modified map $[d(r_l)]^2/[p(0) + p(2r_l)]$. Thus, when there is an inversion symmetry, this modified dirty map should be computed, and the new components should be placed at its symmetric maxima. The strength of the components is given by (10), except for the optional use of the gain factor g.

The generalization to other symmetries is quite simple. In the general case, for every peak at r_l, let there be symmetry related peaks at r_l', r_l'', etc. Then, the new components should be placed at the maxima r_k, r_k', r_k'', etc., of the map

$$\frac{[d(r_l)]^2}{p(0) + p(r_l - r') + p(r_l - r_l'') + \dots}, \tag{13}$$

and the strength of the component should be

$$S = \frac{g d(r_k)}{p(0) + p(r_k - r_k') + p(r_k - r_k'') + \dots} \tag{14}$$

5. Some illustrative simulations

We have carried out calculations on simple one and two dimensional models to test the efficiency of CLEAN in crystallographic problems. We present here some results to illustrate the potential of the method.

Figure 1(a) shows a one-dimensional 'crystal' having three 'atoms' per unit cell. The electron density in the unit cell is of the form

$$\rho(n) = 0.3 \exp[-(58-n)^2/196] + \exp[-(118-n)^2/144]$$

$$+ 0.4 \exp[-(198-n)^2/225]; \quad n = 0, 1, 2, \dots, 256. \tag{15}$$
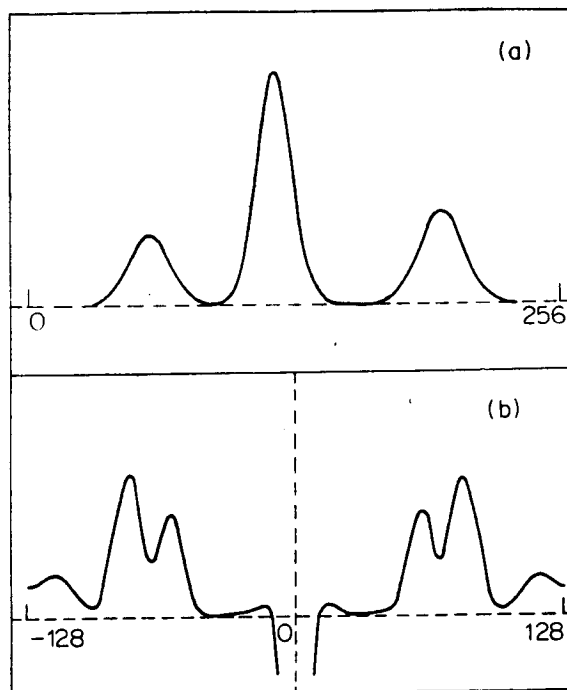
Figure 1. (a) Assumed electron density in a one-dimensional model structure with three atoms (b) Origin-removed sharpened Patterson of this structure.

Figure 1(b) shows the origin-removed sharpened Patterson (the sharpening function used was exp $(n^2/321)$). The structure can be easily solved from this Patterson.

To test CLEAN, we took the above structure and assumed that we have measured only the first eight Fourier coefficients *viz.* $F(0)$ to $F(7)$. When we sharpen this highly truncated data and calculate the origin-removed Patterson, we obtain the map shown in figure 2(a). It is quite clear that the structure cannot be directly solved from this, because of the termination ripple and the loss of resolution. The PSF for the case we are considering is shown in figure 2(b).

Using the dirty map 2(a) and the PSF 2(b), we implemented CLEAN using a gain of 0·5. We then convolved the components with clean peaks of various half-widths. Figures 3(a) to (d) show the results for Gaussians clean peaks of the form exp $(-n^2/\sigma^2)$ with $\sigma = 11$, 8, 6 and 4.8. Since the resolution (full width at half maximum) of the data is 20·6 units, these correspond to resolution enhancements of 1·03, 1·55, 2·06 and 2·58 respectively. The CLEAN-ed Pattersons of figure 3 are seen to be much superior to the dirty Patterson of figure 2(a).

Figure 4 shows a more complicated two-dimensional simulation. Figure 4(a) gives the positions of seven equal atoms in a two-dimensional structure while figure 4(b) shows the peaks in the origin-removed Patterson of this structure (the origin is at the centre of the map). Taking only a restricted set of E's with (h, k) limited to the range -6 to $+6$, one computes the sharpened Patterson shown in figure 4(c). The true positions of the Patterson peaks are also indicated for reference. It will be noticed that clusters of nearby peaks have become merged into single peaks and the isolated peaks cannot be located. There are also many negative regions. Figure 4(d)
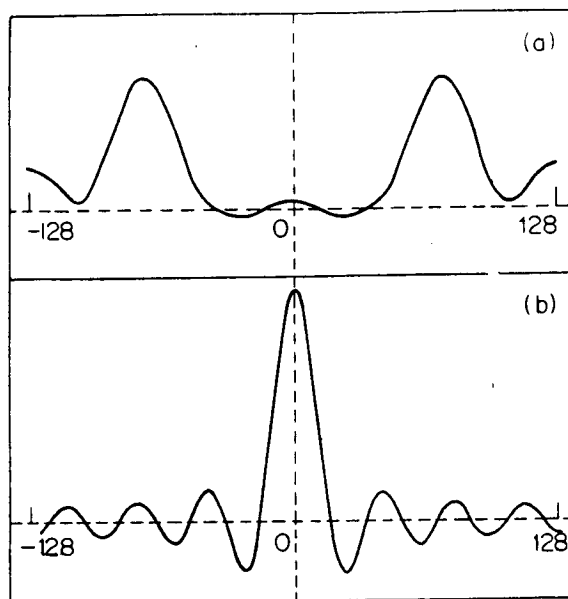
Figure 2. (a) Origin-removed sharpened Patterson of the structure in figure 1 (a) calculated using the first eight structure factors *viz.* $F(0)$, $F(1)$, ..., $F(7)$. (b) The point spread function (PSF) for the case when $F(0)$, $(F(1)$, ..., $F(7)$ alone are included.
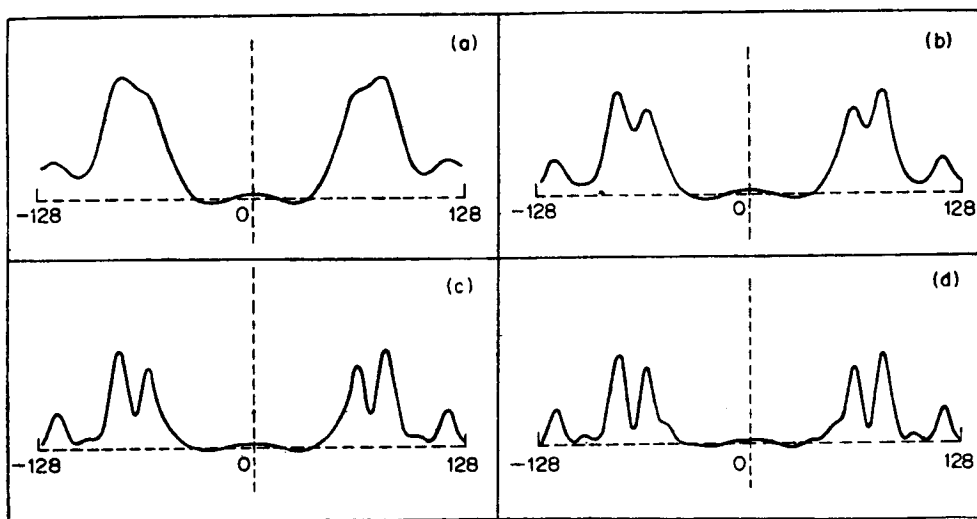

Figure 3. CLEAN-ed Patterson maps obtained from figure 2 (a) using the PSF in figure 2 (b) with a gain factor of 0·5. Clean peaks $\exp(-n^2/\sigma^2)$ were used with (a) $\sigma = 12$ (b) $\sigma = 8$ (c) $\sigma = 6$ and (d) $\sigma = 4.8$.

shows the CLEAN map obtained from figure 3(b) with a gain of 0·5 and a resolution enhancement of 1·70. The improvement is obvious. The negative regions have disappeared and all the single peaks can be clearly identified. Some of the merged peaks have also been separated. Even in those cases where the peaks continue to be merged, the resultant peaks are elongated along the line of separation. It would clearly be much easier to interpret figure 4(d) than 4(c).
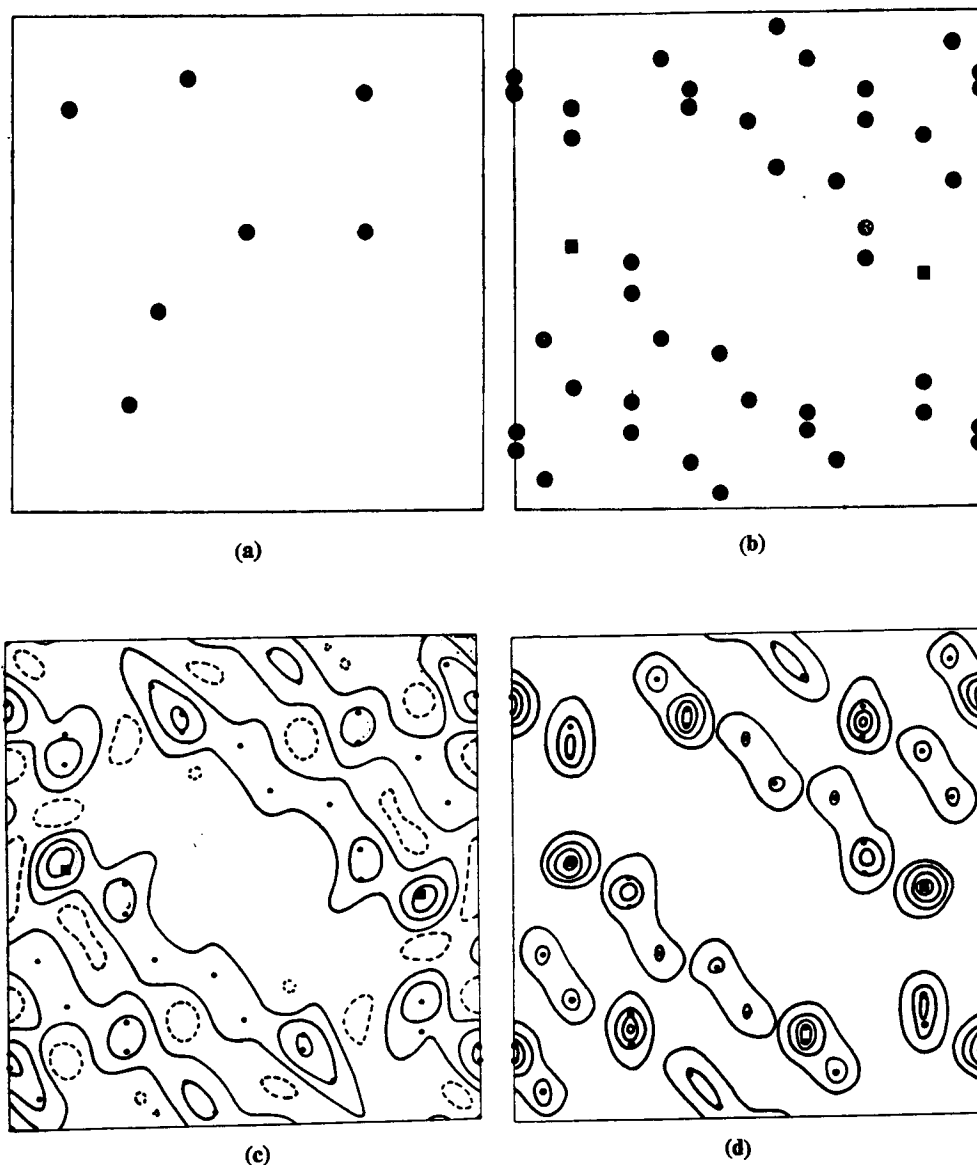
(a)   (b)

(c)   (d)

Figure 4. (a) Model two-dimensional structure with seven equal atoms. (b) Peaks in the origin removed Patterson. The origin is at the centre. Of the 42 peaks in the unit cell, two pairs of peaks overlap to form double peaks. These are indicated by solid squares. (c) Origin removed sharpened Patterson calculated with $|E(h, k)|^2$, $|h|, |k| \leqslant 6$. Negative and positive regions are shown by dashed and solid contours respectively. The negative contours are at -0.05 while successive positive contours are at $0.05, 0.15, 0.25$, etc. On this scale, the single peaks in (b) have height 1 and double peaks height 2. (d) CLEAN-ed Patterson map obtained from (c) with $g = 0.5$ and a Gaussian clean beam corresponding to a resolution improvement of 1.70.

The above simulations show that CLEAN can be very profitably used both to remove ripple as well as to increase resolution. The method therefore seems to be suitable for many crystallographic applications.

Acknowledgements

References

Högbom J A 1974 *Astron. Astrophys. Suppl.* **15** 383
Schwarz U J 1978 *Astron. Astrophys.* **65** 345
Stout G H and Jensen L H 1968 *X-ray structure determination: A practical guide* (London: Macmillan)