Acta Cryst. (1978). A**34**, 353–361

The Recognition of Molecular Fragments in *E* Maps and Electron Density Maps

BY P. MAIN AND S. E. HULL

Department of Physics, University of York, York YO1 5DD, England

(*Received* 10 *October* 1977; *accepted* 22 *November* 1977)

The advent of phase-determination computer programs like *MULTAN* has made the solution of many crystal structures a matter of routine. The next major step towards making the process completely automatic is to program the computer to recognize complete molecules and molecular fragments in *E* maps. Algorithms suitable for the machine interpretation of *E* maps are given. These include procedures for peak search, separation of peaks into bonded clusters, application of stereochemical criteria and comparison of the molecular fragments found with those expected. The algorithms are capable of giving alternative interpretations of the same map. Two methods of comparing molecular fragments found in *E* maps with the expected molecular structures are described. Both have been used successfully and examples of their use are presented. One of the methods can be programmed fairly easily in a completely general way and will be incorporated into *MULTAN*. The recognition of molecular fragments will allow the automatic use of a partial structure as a step in the determination of the complete structure.

Introduction

In recent years, advances in direct methods of phase determination have made the solution of many crystal structures a matter of routine. However, there remain structures which are not easy to solve and in which it may be necessary to look at many different potential solutions before finding the correct one. Sometimes the best *E* maps may only yield part of the structure, possibly incorrectly positioned, and the partial structure must be used as a step towards the complete solution. In both cases, it would be an advantage to have the *E* maps interpreted directly by the computer. This would mean that, at least, stereochemically sensible molecular fragments are detected automatically or, better, the computer compares the molecular fragments found with those expected and informs the user which peaks in the map are likely to correspond to atoms in the structure.

When an *E* map reveals only part of the structure, there exist powerful direct-methods techniques which make use of this information (for example, Karle, 1976; Main, 1976). If structure solution is to be fully automated, even for difficult structures, these techniques must be implemented in a form which requires a minimum of human intervention. A vital step in this process is the automatic interpretation of *E* maps in which not only complete molecules, but also molecular fragments can be recognized.

The problems involved in molecule recognition have been discussed from a theoretical standpoint by Bart & Busetti (1976) and, more practically, by Koch (1974). In this paper we describe two methods which have been used successfully in the interpretation of *E* maps. One of them is already incorporated in a limited form into the current version of the *MULTAN* system of programs (Main, Lessinger, Woolfson, Germain & Declercq, 1977). The other, which is easier to program more generally, will be implemented in the next version. In both methods, the interpretation of *E* maps (or electron density maps) is divided into four distinct stages: (1) peak search, *i.e.* obtain a list of the highest unique peaks in the map; (2) separation of peaks into potentially bonded clusters; (3) application of simple stereochemical criteria to produce potential molecular fragments; (4) comparison of the fragments found in stage 3 with the expected molecular structure. Stages (1) and (2) have already been described by Declercq, Germain, Main & Woolfson (1973) and will not be discussed at length here. Stages (3) and (4) were dealt with by Koch (1974) and it is these that form the main subject matter of this paper.

Peak search

The searching of electron density and *E* maps for maxima is now an operation which is commonly performed on a computer. Normally, the position of a maximum is found by interpolation after fitting a function of an appropriate form by least squares to 19 or 27 points around the maximum (see, for example, Dawson, 1961). The most suitable function for electron density maps is either a Gaussian or a very similar function involving more than one exponential term. For *E* maps, the peak shape is, theoretically, the Fourier transform of a solid sphere, *i.e.* a function of the form:

$$f(x) = \frac{3(\sin x - x\cos x)}{x^3}. \qquad (1)$$

In practice, however, the peak shape deviates considerably from this because of errors in the phases and because only a few hundred reflexions have been used to compute the map. It is, therefore, sufficient to use a quadratic function, which is considerably simpler than (1) above, provided that only the immediate vicinity of the maximum is explored. In addition, a spherically symmetric function may be used, which means the only parameters to be determined are the peak height, position and width. This leads to a function of the form:

$$\rho(x,y,z) = a + cx + dy + ez - \tfrac{1}{2}f(x^2 + y^2 + z^2) \quad (2)$$

where the five parameters a, c, d, e, f are evaluated from a least-squares fit to 19 points. Expressions for the parameters in terms of the density at the 19 points are given in Table 1. The position of the maximum relative to the central point is given by:

$$(u,v,w) = \left(\frac{c}{f}, \frac{d}{f}, \frac{e}{f}\right) \quad (3)$$

and the value of the function at the maximum is:

$$\rho_{\max} = a + \tfrac{1}{2}(cu + dv + ew) \quad (4)$$

where u,v,w are given in (3). For E maps, $\rho_{\max}$ as computed from (4) cannot be used as a reliable measure of the peak height since the peak shape is normally very different from that assumed by (2), resulting in a poor fit. It is better to use the maximum grid value instead.

A comparison of peak positions obtained by fitting the quadratic function (2) with those obtained from the more general function

$$\rho(x,y,z) = \exp(a + cx + dy + ez - \tfrac{1}{2}fx^2 - \tfrac{1}{2}gy^2 - \tfrac{1}{2}hz^2 + pxy + qyz + rzx) \quad (5)$$

has been made. The r.m.s. difference in E-map peak position from that of the refined structure was determined for a small number of structures and was found to be about 0·1 Å using either (2) or (5). The simpler function (2) is, therefore, the one that we currently use.

Table 1. *Expressions for the parameters in equation* (2) *in terms of the density at the* 19 *grid points around the maximum*

b_{ijk} = density at the grid point (i,j,k) where b_{000} is the maximum

i,j,k = $-1,0,+1$ and at least one of i,j,k must be zero

let $P = b_{\bar{1}00} + b_{0\bar{1}0} + b_{00\bar{1}} + b_{001} + b_{010} + b_{100}$

let $Q = b_{\bar{1}\bar{1}0} + b_{\bar{1}0\bar{1}} + b_{\bar{1}01} + b_{\bar{1}10} + b_{0\bar{1}\bar{1}} + b_{0\bar{1}1} + b_{01\bar{1}} + b_{011} + b_{1\bar{1}0} + b_{10\bar{1}} + b_{101} + b_{110}$

then $21a = 9b_{000} + 4P - Q$

$10c = b_{1\bar{1}0} + b_{10\bar{1}} + b_{100} + b_{101} + b_{110} - (b_{\bar{1}\bar{1}0} + b_{\bar{1}0\bar{1}} + b_{\bar{1}00} + b_{\bar{1}01} + b_{\bar{1}10})$

$10d = b_{\bar{1}10} + b_{01\bar{1}} + b_{010} + b_{011} + b_{110} - (b_{\bar{1}\bar{1}0} + b_{0\bar{1}\bar{1}} + b_{0\bar{1}0} + b_{0\bar{1}1} + b_{1\bar{1}0})$

$10e = b_{\bar{1}01} + b_{0\bar{1}1} + b_{001} + b_{011} + b_{101} - (b_{\bar{1}0\bar{1}} + b_{0\bar{1}\bar{1}} + b_{00\bar{1}} + b_{01\bar{1}} + b_{10\bar{1}})$

$63f = 30b_{000} + 11P - 8Q$

Cluster formation

After obtaining a list of the unique peaks in the map, the next stage in the interpretation is to separate the peaks into potentially bonded clusters. In a cluster, each peak is within chemical bonding distance of at least one other peak in the same cluster. For convenience the positions of the peaks in each cluster are projected on the least-squares plane and plotted on the lineprinter to form a 'picture' of the molecule. Two further projections can also be output if desired. These are on the plane orthogonal to the least-squares and most-squares planes and on the most-squares plane.

It may happen that either spurious peaks join the cluster to itself by space-group symmetry or there exists a bonded chain of atoms running through the whole crystal. For ease of plotting the peaks and subsequent interpretations, the 'monomeric' part of the 'polymeric' cluster is built up by the following simple algorithm.

(1) Start the cluster with the highest peak not already in a cluster.

(2) Take the highest peak in the cluster not already considered.

(3) Add to the cluster all peaks within bonding distance of the peak under consideration.

(4) Repeat from (2) till no more unique peaks are added to the cluster.

This algorithm ensures that the cluster is terminated at the weakest peaks in the map and the highest peaks tend to be in the middle of the cluster.

The number of unique peaks normally considered is about 23% more than the number of unique atoms to be found. However, if two peaks in different clusters are within 2·8 Å of each other, the program assumes the clusters belong to the same molecule and automatically searches for more peaks. In this case, the number of peaks is increased to 50% more than the number of atoms. Ideally, this will amalgamate the clusters so that the complete molecule can be found in a single cluster of peaks instead of being split between two or more. This makes recognition of the molecule much easier.

Application of chemical rules

Koch (1974) described an algorithm for applying the 'rules' of organic chemistry to a set of E-map peaks. The process involves the systematic elimination of peaks until the rules are not violated. This always produces a unique interpretation for each separate cluster of peaks in the E map, but the interpretations

tend to be rather limited. Often, small molecular fragments are produced in cases where chemists can easily identify much larger ones. In order to produce a more exhaustive search for molecular fragments, one of us (PM) has devised an algorithm which is capable of producing several different interpretations of the same cluster of peaks and is generally less restrictive than Koch's method. Rather than eliminate peaks in a unique manner, the new algorithm builds up molecular fragments according to simple stereochemical criteria. Different peaks may be assumed to be spurious in order to fulfil the criteria, thus allowing several interpretations of the cluster.

The steps in the process are as follows.

(1) Start the interpretation with the highest peak not already used in an interpretation; this peak is the molecular fragment to which other peaks are added.

(2) Find the highest peak in the fragment not already considered in this step; call this peak *B*.

(3) Eliminate all peaks which are too close to *B*, *e.g.* within 1·1 Å.

(4) Find all peaks within bonding distance of *B*, *e.g.* within 1·95 Å; take these peaks in pairs, say peaks *A* and *C*, and calculate all bond angles $\angle ABC$.

(5) Eliminate peaks for which $\angle ABC$ is unacceptable, *e.g.* $< 85°$ or $> 145°$, in the following order: (i) if *A* and *C* are already in the fragment, eliminate *B* (in this case, go directly to stage 7); (ii) if *A* (or *C*) is already in the fragment, eliminate *C* (or *A*); (iii) if neither *A* nor *C* is in the fragment, eliminate the weaker of *A* and *C*.

(6) Add all remaining peaks within bonding distance of *B* to the fragment.

(7) Repeat from (2) till no more peaks are added to the fragment.

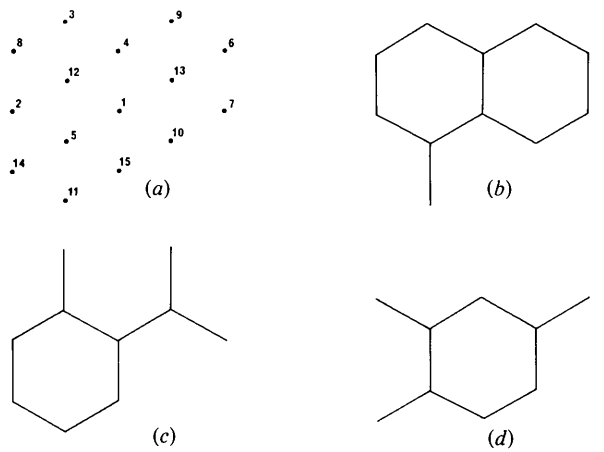


Fig. 1. (*a*) A cluster of peaks on a hexagonal lattice which will allow several interpretations. (*b*), (*c*) and (*d*) Three possible molecular structures which can be fitted to the peaks in (*a*).

(8) If there remain any peaks which are neither in a fragment nor eliminated, choose the highest as the start of a new fragment and go back to stage (2).

(9) If any peaks have not yet been used in an interpretation, *i.e.* accepted as part of a molecular fragment, reinstate all eliminated peaks and go back to stage (1) to search for a new interpretation.

Note that the suggested limits on bond lengths and angles rule out all three and most four-membered rings, triple bonds, coordinations greater than four and genuine long bonds (>1·95 Å). However, they are suitable for the vast majority of light-atom organic compounds and relaxation of the limits would result in too many possible interpretations for the same set of peaks. Stage (8) in the algorithm allows several molecular fragments to be built out of one cluster of peaks while stage (9) allows several different interpretations of the same cluster.

To illustrate the scope and limitations of the algorithm for finding different interpretations, a cluster of peaks is shown in Fig. 1(*a*). The peaks are numbered in order of peak height, peak 1 being the highest. Three possible interpretations in terms of molecular structure are shown in Fig. 1(*b*), (*c*) and (*d*). Following the algorithm, the first interpretation starts at peak 1. Peak 4 is added to the fragment since this is the highest peak within bonding distance of 1. Peaks 12, 13 and 15 are eliminated since they form disallowed bond angles with the 1–4 bond. The remaining peaks around 1, *i.e.* 5 and 10, are then added to the fragment. The largest fragment peak not yet considered is now number 4. Bond lengths and angles around peak 4 are considered, resulting in peaks 3 and 9 being added to the fragment. Following the algorithm through to the end gives the interpretation shown in Fig. 1(*b*). The highest peak not used by this molecule is number 12, so a second interpretation is started at this peak. Using the same logic as before, peaks 1 then 13, 15, 2 and 3 are added to the fragment and, eventually, the interpretation shown in Fig. 1(*c*) emerges. Together, the two interpretations which have been obtained make use of all the peaks in the cluster. The algorithm therefore terminates and never finds the third interpretation shown in Fig. 1(*d*).

Clearly, the search for different interpretations is not exhaustive, but our experience to date indicates that this is not a disadvantage. The algorithm ensures that the first interpretation of the map uses as many of the highest peaks as possible and subsequent interpretations must use peaks of decreasing height. The highest peak in the 'missing' interpretation in Fig. 1(*d*) is fourth in ranking order and maps in which the highest three peaks are spurious are unlikely to yield the structure. However, a small change in the algorithm would make it more exhaustive if this were found necessary. Instead of starting a new interpretation at a peak which has not yet been used (stage 1),

a new start could be made at a bond (*i.e.* a pair of peaks) which had so far not been used. With this change, the interpretation in Fig. 1(*d*) would be found. We have no practical experience of this modified algorithm since it has not yet been programmed.

The overall behaviour of the algorithm is similar to that of a human interpreter in that it produces a single, correct interpretation when the answer is obvious and produces a large number of possible interpretations when faced with a rather 'messy' *E* map. In this latter case, the information output by the computer on possible bonding patterns is a valuable aid towards deciding which peaks in the map constitute part of the real structure. However, the next sections describe methods by which the computer can make these choices for itself.

Comparison of molecular fragments

The problems of comparing a molecular fragment with the expected structure have been well described by Koch (1974) and by Bart & Busetti (1976). In general, a molecular fragment, such as that obtained from an *E* map, as already described, may contain spurious atoms and a number of genuine atoms may be missing. The purpose of the comparison of the fragment with the expected structure is to detect the spurious atoms and identify the genuine atoms that are present.

An exhaustive comparison is out of the question because of the huge number of ways in which atom-by-atom correspondence can be achieved. If there are m peaks in the fragment and n atoms in the molecule, r of which are missing from the fragment, then the total number of comparisons which can possibly be made is

$$N = \sum_{r=0}^{n} {}^{n}C_{r}\,{}^{m}C_{n-r}. \qquad (6)$$

This formula assumes that r is unknown and so all possible values must be tried. Even in the trivial case of $m = 10$ and $n = 8$, this amounts to 43758 comparisons. Clearly, a much more intelligent approach than this is required. A human being would not attempt an interpretation in this way since the vast majority of these 'structures' do not make chemical sense. An approach which retains the chemical integrity of both fragments would therefore be an advantage.

Neither position nor orientation of the fragment can be used to identify it. Also, information on atomic type will not be available from the *E* map. In general, the stereochemistry of the expected structure cannot be used for identification because it will not always be completely known. Also, the stereochemistry of the *E*-map fragment will be inaccurate. The use of stereochemical criteria as described in the previous section represents the extent to which this information is used at this stage.

The information which is most readily available, and by which most people will recognize the structure, is the molecular connectivity. We have devised two ways of using connectivity information to interpret molecular fragments and these are described in the next two sections. The comparison of the molecular structures is thus seen to be topological rather than geometrical.

Direct comparison

The first method of identifying molecular fragments is by direct comparison of the two connectivities. It is most easily applied when both fragments contain ring systems. A ring system is defined as a fragment in which all the points (atoms or peaks) are connected to at least two other points. The method attempts to find a ring system which is common to both fragments. If such a ring system is found, corresponding atoms are identified, then the side groups are added in so far as the *E*-map peaks correspond to the expected structure. The stages in the comparison are as follows.

(1) Reduce both fragments to ring systems by eliminating all points with a connectivity of unity until no such points remain. This should eliminate many of the spurious *E*-map peaks and also eliminate atoms towards the periphery of the molecule which may not be observed because of high temperature factors.

(2) Compare the number of points in the two ring systems. If they contain unequal numbers of points, go to stage (4).

(3) Calculate the canonical description of each ring system (Morgan, 1965) and compare them. If the two structures have identical canonical descriptions, go to stage (5).

(4) Delete one or more ring atoms in one of the fragments, leaving a ring system which has not yet been examined at stage (3). Go back to stage (1). If no new ring system is possible, the two fragments cannot be matched.

(5) Identify corresponding atoms in the two ring systems. This is easily done since, if the ring system is asymmetric, the canonical description associates a unique number with each atom. Clearly, those atoms in different fragments with the same number in the canonical descriptions must correspond.

(6) Build up the side groups on the rings as far as *E*-map peaks and atoms in the expected molecule allow.

The canonical description algorithm, referred to in stage (3), produces a unique set of numbers for a given connected structure, each number corresponding to a point in the structure. It is equivalent to finding the principal eigenvector of the connectivity matrix $\mathbf{A} = (a_{ij})$, where a_{ij} is unity if points i and j are connected and zero otherwise. In practice, the

canonical description is calculated by performing a number of cycles of the operation

$$\mathbf{C}_{n+1} = \mathbf{A}\mathbf{C}_n \tag{7}$$

where $\mathbf{C}_n$ is the nth approximation to the canonical description and $\mathbf{C}_0$ is the vector containing the number of connections to each point.

In the original form of the algorithm, the matrix multiplication was continued until the number of different values of the elements in the vector $\mathbf{C}_n$ ceased to increase. This does not always give the maximum possible number of unique elements in $\mathbf{C}_n$, so we now make the number of cycles equal to the number of points in the fragment. Since the atoms in the two fragments will not necessarily be considered in the same order, the canonical descriptions are best compared by first arranging the elements in numerical order and then comparing corresponding elements.

The systematic alteration of the ring systems in stage (4) is easily arranged and is best explained by taking an actual example. Fig. 2(*a*) shows a ring system which might be contained in one of the fragments. By eliminating a point of connectivity two in each of the four fused rings in turn, the four different ring systems which remain are shown in Fig. 2(*b*). In each case, the point eliminated to produce the new ring system is present in the previous ring systems. The 'molecules' in Fig. 2(*b*) can, therefore, be generated very quickly and used in the comparison with the other fragment. Similarly, the ring systems shown in Fig. 2(*c*) are all generated by deleting a single point of connectivity three from the original ring system of Fig. 2(*a*). Note that the middle two are topologically the same. Combinations of two or more points can then be eliminated from the original ring system to produce other ring systems for comparison. In this way, one can generate all possible ring systems contained in the original in a very short time.

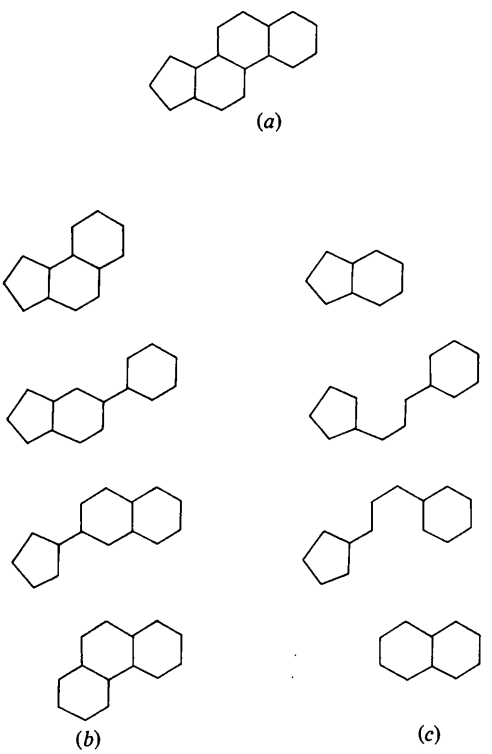


Fig. 2. (*a*) Example of a ring system to be found in a molecule. (*b*) Ring systems generated from (*a*) by the elimination of a point of connectivity 2. (*c*) Ring systems generated from (*a*) by the elimination of a point of connectivity 3.

Once the ring atoms have been identified in stage (5), the side groups can be identified in stage (6) using logic very similar to that described by Sussenguth (1965). Fig. 3(*a*) shows part of a molecule which is to be matched to the set of *E*-map peaks in Fig. 3(*b*). Assume atoms 1, 2 and 3 belong to a ring system and these have already been identified with peaks *h*, *b* and *a* respectively at stage (5). This may be written

$$1 \equiv h; \quad 2 \equiv b; \quad 3 \equiv a. \tag{8}$$

The atoms bonded to 2 are 1, 3 and 4 and these must correspond to the peaks bonded to *b* which are *a*, *c* and *h*. This may be written

$$\{1,3,4\} \equiv \{a,c,h\}. \tag{9}$$

However, we know that $1 \equiv h$ and $3 \equiv a$. We therefore conclude that $4 \equiv c$.

Examination of the atoms bonded to 4 and the peaks bonded to *c* gives

$$\{2,5,6\} \equiv \{b,d,g\}. \tag{10}$$

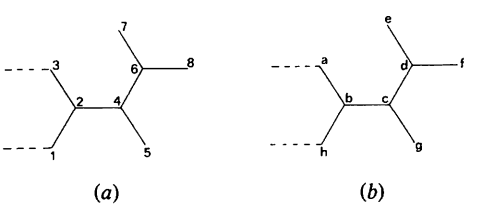
since $2 \equiv b$, we obtain

$$\{5,6\} \equiv \{d,g\}. \tag{11}$$

To resolve the ambiguity, we can examine the connectivities of the atoms and peaks in (11), *i.e.* to how many other atoms or peaks they are bonded. The connectivity of 5 is 1 and that of 6 is 3 while the connectivity of *d* is 3 and that of *g* is 1. It is clear, then, that $5 \equiv g$ and $6 \equiv d$.

In attempting to match the remaining atoms and peaks, we obtain

$$\{7,8\} \equiv \{e,f\}. \tag{12}$$



(*a*) (*b*)

Fig. 3. (*a*) Naming of atoms in molecule which is to be identified with the *E*-map fragment in (*b*). (*b*) Naming of peaks in *E*-map fragment.

The implied ambiguity in (12) does not, in fact, exist. This is because atom types are not identified in the *E* map and so it is immaterial whether atom 7 is identified with peak *e* or *f*. Any peak with a connectivity of unity, bonded to peak *A*, say, can always be identified with an atom of unit connectivity, bonded to atom 1, say, provided peak *A* corresponds to atom 1. Thus, the complete side group of the molecule has been matched to the set of peaks.

In practice, the molecular fragment obtained from the *E* map may contain spurious peaks, or some peaks may be missing. This means that the use of connectivities to resolve ambiguities as in (11) above is not always possible. Even if the particular side group matches the molecule exactly, an ambiguity can arise which cannot be dealt with in the manner described. In both cases, the ambiguity can be resolved by referring to the canonical descriptions of the two structures. It has been observed that points in similar positions in two similar structures are usually in the same position in the ranking order of their canonical weights. Therefore, the canonical descriptions of both the fragment and the molecule are computed and the peaks and atoms in question are listed in decreasing order of their canonical weights. The first peak is identified with the first atom and the remaining peaks and atoms are matched using the logic just described. Any further ambiguities are resolved by recourse to the canonical descriptions. This does not guarantee that ambiguities are always correctly resolved, but in most cases they should be. The same technique is used to match ring systems at stage (5) when symmetry prevents a unique peak–atom correspondence to be made. (Note that a hexagon of points can be matched to another hexagon in 12 different ways.)

This whole method of comparing molecular fragments has not yet been programmed in the way described. We have only implemented a limited version of it, partly because the program was more complicated than was justified by the current state of development of the *MULTAN* system, and also because of the success of the trial and error method described in the next section. However, it performs comparisons successfully and very efficiently, taking only a few tenths of a second on a PDP10 for a structure containing 40–50 atoms. Its main disadvantage is that it depends upon the fragments containing ring systems. When no ring system is present, a modification of the method must be used.

In the absence of rings, an initial match may be attempted between the 'backbones' of the two fragments. The backbone is defined as the longest path that can be taken from one terminal atom to another which produces as many side chains as possible. For example, Fig. 4(*a*) shows a molecule without a ring system, the longest path of which consists of six atoms, either from atom *a* to *l* or from *a* to *h*. (Note that *a* and *c* are indistinguishable, as are *j* and *l*.) Fig. 4(*b*) shows a topological equivalent of the molecule which emphasizes the *a*–*l* path and it can be seen that this has four side chains. Fig. 4(*c*) shows that the *a*–*h* path has three side chains. Therefore, the backbone in this case is taken as *a*–*l*. The backbone may be used in a similar role to that of the ring system when comparing non-ring fragments, but this has not been tried. It may, indeed, be unnecessary, since the method described in the next section is sufficiently general to deal with any type of molecule in a single procedure.

Trial and error comparison

Although the direct-comparison method is very efficient, it requires different procedures to deal with ring and non-ring fragments, which results in a fairly lengthy program. In addition, the use of the canonical descriptions to deal with ambiguities is not completely reliable and, to be safe, several possible fits of one fragment on the other must be tried. This leads us directly to the trial and error comparison method. If trial and error must be used at all, can it be used for the complete matching process?

It has already been pointed out that an exhaustive comparison of the fragments is quite impractical. On the other hand, we know from observation that points in similar positions in similar fragments are usually in similar positions in the ranking order of their canonical weights. In addition, those points near the centre of the fragment tend to have higher weights than those towards the periphery. Therefore, if the points in two different fragments are ordered on decreasing canonical weight, the points near the centre of each fragment will appear near the top of the appropriate list. Moreover, corresponding points in the two fragments will be in similar positions in the two lists. This means that in searching for a point in one structure corresponding to a particular point in the other structure, the most likely points to produce a match can be identified. Therefore, there is a high probability

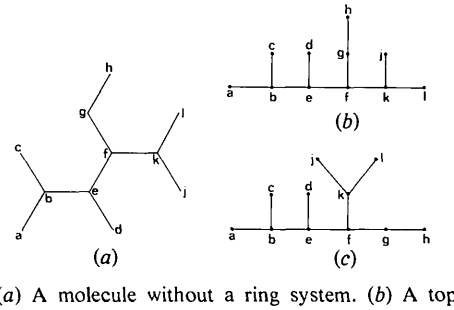


Fig. 4. (*a*) A molecule without a ring system. (*b*) A topological equivalent of (*a*) emphasizing the path from atom *a* to atom *l*. (*c*) Similar to (*b*) emphasizing the path from *a* to *h*.

that the correct point will be found early in the search. The search is restricted still further by insisting that both fragments keep their chemical integrity throughout. This is the basis of the trial and error comparison method.

The method is effectively an exhaustive search, with constraints, for corresponding points in the two fragments with a good change of success very early in the search. However, it was found necessary to develop an empirical device to speed up the search still more. The resulting algorithm does not necessarily produce the largest possible number of matched atoms, but the fit is nearly always acceptable, even in the most difficult cases. The stages in the procedure are as follows.

(1) Sort the points in each fragment according to the following rules: (*a*) all ring points, if any, come before all non-ring (*i.e.* chain) points; (*b*) all points, except the first, must be connected to at least one other point higher in the list; (*c*) points are placed in order of decreasing canonical weight, subject to the conditions (*a*) and (*b*) above. This produces two lists, one for each fragment.

(2) Weight the ith point in the list of E-map peaks by

$$W_i = R_i + 10\,000\left(\frac{c_i}{c_{\max}}\right)^2 \qquad (13)$$

where $R_i = 20\,000$ for ring points and zero for all others; c_i is the canonical weight for this point and $c_{\max}$ is the maximum canonical weight for any point in the fragment. The algorithm searches for the match between the fragments which maximizes the sum of the weights over those peaks to which atoms have been assigned.

(3) Assume the first atom (in the expected molecule) corresponds to the first peak (in the E-map fragment).

(4) Take the next peak in the list and find which other peaks higher in the list it is bonded to. If there are no further peaks in the list, go to stage (7).

(5) Test each atom, starting from the top of the atom list, to find one which is not already assigned to a peak and which is bonded to the atoms corresponding to the peaks found in stage (4). If no such atom can be found, go to stage (7).

(6) Assign the atom found in stage (5) to the peak in stage (4) and go back to stage (4).

(7) Determine Σw_i for the peaks currently matched and compare this with the maximum sum of weights, $\Sigma_{\max}$, obtained during the whole search so far. If $\Sigma_{\max}$ can still be exceeded, even if the current peak is ignored, assume the peak is spurious and go back to stage (4). If $\Sigma_{\max}$ cannot be exceeded, even if all the remaining peaks are matched with atoms, go back to stage (4) to consider the *previous* peak in the list and try an alternative assignment for it at stage (5). If the current peak is at the top of the list, the matching process is complete.

When the algorithm terminates, the peak–atom assignment corresponding to $\Sigma_{\max}$ is accepted as the best match. Note that the algorithm demands that peak 1 always be matched. This condition will be relaxed in future if experience shows this to be desirable.

The speed of the algorithm is critically dependent upon the weighting scheme given in (13). The weighting scheme is purely empirical and has been chosen from a number of other possibilities because it optimizes the computing time over a wide range of problems. The terms it contains ensure that peaks in the ring system, if any, are considered first. Also, peaks near the centre of the fragment take precedence over those towards the periphery.

This method of comparing fragments obviously cannot be as efficient as the direct comparison described in the previous section. However, computing times are usually about one or two seconds on a PDP10 for a molecule of 40–50 atoms, which is quite acceptable. The main advantages over the previous method are that the program is quite small and relatively uncomplicated and, more importantly, it is completely general and will match fragments whether they contain rings or not.

Practical results

To illustrate the behaviour of this method of comparing molecular fragments, we will use the structure of ergocalciferol (Hull, Leban, Main, White & Woolfson,

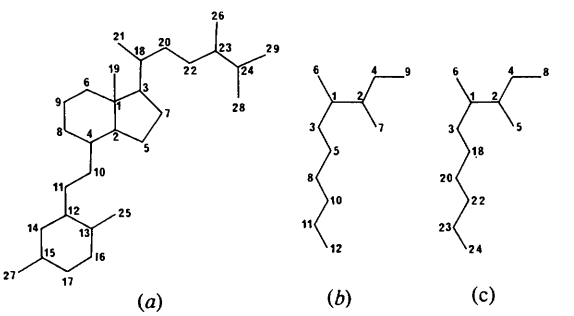


Fig. 5. Ergocalciferol. (*a*) The expected molecular connectivity. (*b*) The fragment connectivity. (*c*) The results of automatically matching the molecule to the fragment.

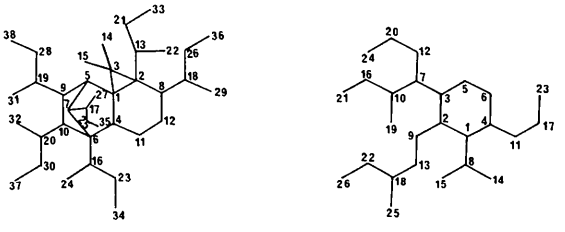
Table 2. *Peak weights for the fragment in Fig.* 5(*b*)

Peak	Weight	Peak	Weight
1	10000	7	1943
2	9299	8	1485
3	5403	9	734
4	3186	10	721
5	2719	11	305
6	2320	12	74

1976). The expected molecular topology is shown in Fig. 5(*a*) and a molecular fragment obtained from an *E* map is shown in Fig. 5(*b*). The numbering in Fig. 5 gives the order of the peaks and atoms after stage (1) of the above algorithm. Table 2 gives the weights of the *E*-map peaks as calculated by (13). The molecule/fragment produced by the algorithm is shown in Fig. 5(*c*), where the numbers give the corresponding atoms of the molecule in Fig. 5(*a*). Since the whole fragment has been topologically matched to the molecule, the algorithm terminates here, even though this is only one of several possible matches. Most of the peaks in the fragment do, in fact, correspond to atoms. The structure was completed using the known orientation of this fragment in the method of Main (1976) to estimate the weights and phases of Σ_2 relationships.

A more complicated example is provided by the *RR* compound originally solved by Declercq, Germain & Henke (1973), which has the connectivity shown in Fig. 6(*a*). A fragment from a poorly resolved *E* map is shown in Fig. 6(*b*) and the weights are listed in Table 3. Fig. 7 shows successive attempts by the program to match the fragment with the molecule. Fig. 7(*a*) shows the first two interpretations. The first of these proceeds smoothly until peak 25 is matched to atom 30; the program then goes back to peak 22 and finds an alternative match (atom 30 instead of atom 32), thus enabling one additional peak (number 26) to be matched. To increase Σw_i further, it is necessary to go back and rematch peak 13 to atom 16 instead of atom 10; this results in the interpretation shown in Fig. 7(*b*). The next step taken by the program is to rematch the ring peaks to atoms in the other six-membered ring of the molecule. In order to do this, it is necessary to go back as far as peak 3 and match it with atom 6 instead of 11. This eventually leads to the interpretations shown in Fig. 7(*c*). The next pathway followed by the program, shown in Fig. 7(*d*), matches the six-membered ring as in Fig. 7(*a*) except for a rotation of 180° about the line joining atoms 1 and 12. Fig. 7(*e*) shows a further match obtained by the program when the six-membered ring of Fig. 7(*a*) is rotated by 60° in the plane of the ring; this is the first interpretation that does not have peak 1 matched to atom 1. The match finally accepted is shown in Fig. 7(*f*) and corresponds to an in-plane rotation of the ring of Fig. 7(*c*) by 120°. All molecule/fragment matches found by the program after this have a lower value of Σw_i.

The final topological fit matches only 19 peaks. This is less than the maximum obtained during the whole



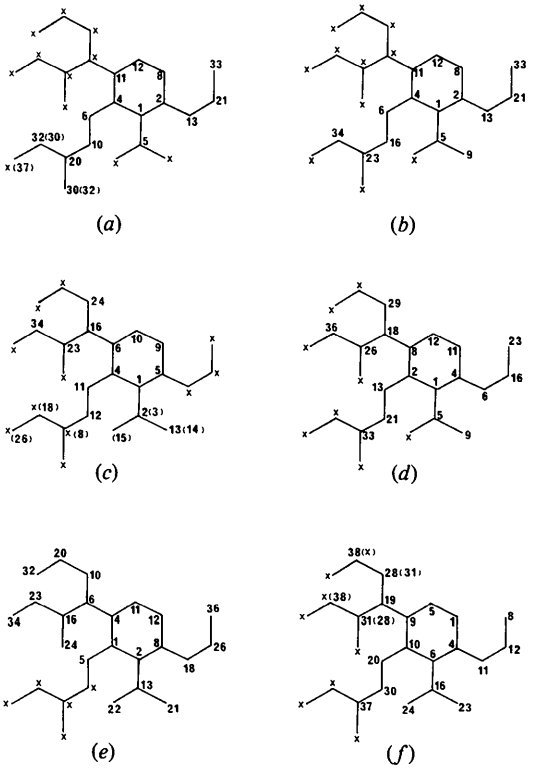
(*a*) (*b*)

Fig. 6. *RR* compound. (*a*) The expected molecular connectivity. (*b*) The fragment connectivity.

Table 3. *Peak weights for the fragment in Fig.* 6(*b*)

Peak	Weight	Peak	Weight
1*	30000	14	613
2*	29693	15	613
3*	29815	16	518
4*	26100	17	373
5*	23810	18	366
6*	23551	19	360
7	5269	20	325
8	3369	21	93
9	2924	22	80
10	2357	23	68
11	1726	24	59
12	1482	25	56
13	812	26	15

* Ring peak.



(*a*) (*b*) (*c*) (*d*) (*e*) (*f*)

Fig. 7. Successive stages in the automatic matching of the molecule to the fragment. In (*a*), (*c*) and (*f*) the figures in brackets refer to a second similar, but slightly better matching. (*a*) Σw_i = 172 743 (172 758), 15 atoms (16 atoms), (*b*) Σw_i = 173 300, 15 atoms, (*c*) Σw_i = 180 313 (181 387), 14 atoms (18 atoms), (*d*) Σw_i = 182 846, 18 atoms, (*e*) Σw_i = 183 118, 21 atoms, (*f*) Σw_i = 183 266 (183 459), 19 atoms (19 atoms).

process which is 21 peaks in Fig. 7(*e*). On purely topological grounds, this is reasonable, as more peaks close to the fragment centre are matched in Fig. 7(*f*) than in Fig. 7(*e*). Also, it is not obvious that the most suitable fit between two different fragments is the one that matches most individual peaks, regardless of their position in the fragment. This entire process took only just over one second of computer time on a PDP10. It was necessary to try only a few thousand potential matches, even though the example is complicated by the presence of many spurious peaks and atoms which cannot be matched and would not be easy for a human being to interpret. In fact all 19 matched peaks correspond to atoms in the structure.

An obvious development of the matching procedure would take the geometries of the molecule and fragment into consideration. Fig. 5 provides an example where this might be desirable. Having assigned peaks to atoms 1 and 2, the atoms 3, 4, 5 and 6 bear a specific geometric relation to each other and, for example, it should be possible for the peaks assigned to atoms 6, 1, 2 and 4 to be part of a six-membered ring. Any program taking such geometric facts into account would be much more complicated than the one described here. It would also be more difficult to use because molecular geometry is not usually known to the same precision as molecular topology.

Discussion

The current version of the *MULTAN* program system (Main *et al.*, 1977) includes the *E*-map interpretation routines described above and a limited version of the direct fragment comparison method. The application of chemical rules to the interpretation of *E* maps has been found to be very useful, especially when the map contains a large number of spurious peaks. However, the identification of the peaks in terms of the expected molecule is at too primitive a state of development to be of real practical value at the moment.

Our tests show that the trial and error comparison of fragments works satisfactorily. It is much easier to program in a completely general way than the direct comparison and so will be incorporated into the next version of *MULTAN*. The next stage of development of this technique will be to use the expected molecular structure actively in the interpretation of the *E* map. At the moment, it plays a passive role and is merely used as a check on the structure obtained. Active use of the expected structure will make it possible to deal with a wider range of structures such as those containing features disallowed by the chemical rules as they are currently applied.

Since the trial and error comparison method is completely general, it will be capable of dealing with any kind of connected structure. A cage structure, for example, can be handled in the same way as a ring structure. If the expected molecule is not known exactly but there are a small number of possibilities, each of these expected molecules in turn can be compared with the structure found. In a favourable case this will identify the correct possibility unambiguously. If the expected molecule is in error the algorithm will still do its best to match it with the structure found. In this case, presumably, a small part of the two molecules will correspond, but the computer will not be sufficiently intelligent to inform the user that it believes the *E* map rather than the chemical information supplied.

Interpretation of *E* maps and recognition of the structures they contain is something a human being can do much better than a computer. Possibly, this may always be the case. However, it is clear that a computer can now perform simple recognition tasks without human intervention. This makes it possible to use partial structural information automatically as a step towards the determination of the complete structure.

References

BART, J. C. J. & BUSETTI, A. (1976). *Acta Cryst.* A**32**, 927–933.

DAWSON, B. (1961). *Acta Cryst.* **14**, 999–1000.

DECLERCQ, J.-P., GERMAIN, G. & HENKE, H. (1973). *Cryst. Struct. Commun.* **2**, 405–409.

DECLERCQ, J.-P., GERMAIN, G., MAIN, P. & WOOLFSON, M. M. (1973). *Acta Cryst.* A**29**, 231–234.

HULL, S. E., LEBAN, I., MAIN, P., WHITE, P. S. & WOOLFSON, M. M. (1976). *Acta Cryst.* A**32**, 538–550.

KARLE, J. (1976). *Crystallographic Computing Techniques*, edited by F. R. AHMED, K. HUML & B. SEDLÁČEK, pp. 155–164. Copenhagen: Munksgaard.

KOCH, M. H. J. (1974). *Acta Cryst.* A**30**, 67–70.

MAIN, P. (1976). *Crystallographic Computing Techniques*, edited by F. R. AHMED, K. HUML & B. SEDLÁČEK, pp. 97–105. Copenhagen: Munksgaard.

MAIN, P., LESSINGER, L., WOOLFSON, M. M., GERMAIN, G. & DECLERCQ, J.-P. (1977). *MULTAN 77. A System of Computer Programs for the Automatic Solution of Crystal Structures from X-ray Diffraction Data.* Univ. York, England, and Louvain, Belgium.

MORGAN, H. L. (1965). *J. Chem. Doc.* **5**, 107–113.

SUSSENGUTH, E. H. (1965). *J. Chem. Doc.* **5**, 36–43.