that (i) the spherically symmetric component of the molecular core charge is more diffuse than for the isolated atom, (ii) the molecular core charge has a dominant dipolar polarization which is clearly evident in the scattering amplitudes and phases, and (iii) a quadrupolar density function, with a much smaller electron population than the dipolar term, gives rise to a small part of the structure in the core molecular density scattering factor. We have emphasized that these same core deformation terms are also present in products of the valence molecular orbitals and that the core molecular orbital product alone accounts for about one-half to two-thirds of the total of such features. The deformation terms have an amplitude of scattering which is no more than 1 % of the total molecular scattering for the range of $\sin\theta/\lambda$ studied in this work ($0 \leq \sin\theta/\lambda \leq 1{\cdot}5$ Å^{-1}). Nonetheless, these density deformations make appreciable contributions to certain physical and molecular properties (in particular electric fields). Quadrupole core deformations can, to some extent, be absorbed by a valence scattering model; dipolar deformation will not be accommodated by this model. We have argued that because of the increased effective site symmetry of a multiply bonded atom, most core deformation terms for it will be smaller than for an atom in a diatomic, whereas for a terminally bonded atom, such as an oxygen in a carbonyl or amide group, they will probably be about the same. We therefore mildly suggest that the density functions in Table 1 be tried for the construction of a dipole generalized X-ray scattering factor in the charge density analysis of a terminally bonded atom.

This research has been supported by NSF Grant GP-22729. Generous support by the Alfred P. Sloan Foundation is gratefully acknowledged. Grants of computer time from the Carnegie–Mellon University Computation Center and from the Mellon Institute NMR Facility for Biochemical research (NIH Grant RR00292) are appreciated.

References

Bader, R. F. W. & Bandrauk, A. D. (1968). *J. Chem. Phys.* **49**, 1653–1665.
Bader, R. F. W., Keaveny, I. & Cade, P. E. (1967). *J. Chem. Phys.* **47**, 3381–3402.
Bentley, J. & Stewart, R. F. (1971). Abstr. ACA, Winter Meet. Paper H11.
Bentley, J. & Stewart, R. F. (1973). *J. Comput. Phys.* **11**, 127–145.
Cade, P. E. & Huo, W. M. (1967). *J. Chem. Phys.* **47**, 614–648.
Cade, P. E., Sales, K. D. & Wahl, A. C. (1966). *J. Chem. Phys.* **44**, 1973–2003.
Clementi, E. (1965). *I.B.M. J. Res. Devel.* **9**, 2 (Suppl.).
Coppens, P. (1971). *Acta Cryst.* B**27**, 1931–1938.
Edmiston, C. & Ruedenberg, K. (1963). *Rev. Mod. Phys.* **35**, 457–465.
Fock, V. (1930). *Z. Phys.* **61**, 126–148.
Groenewegen, P. P. M., Zeevalkink, J. & Feil, D. (1971). *Acta Cryst.* A**27**, 487–491.
Hall, G. G. (1961). *Phil. Mag.* **6**, 249–258.
Hehre, W. J., Ditchfield, R., Stewart, R. F. & Pople, J. A. (1970). *J. Chem. Phys.* **52**, 2769–2773.
Huo, W. M. (1965). *J. Chem. Phys.* **43**, 624–647.
Hurley, A. C. (1954). *Proc. Roy. Soc.* A**226**, 179–192.
Koopmans, T. (1934). *Physica,* **1**, 104–113.
Niessen, W. von (1972*a*). *J. Chem. Phys.* **56**, 4290–4297.
Niessen, W. von (1972*b*). *Theor. Chim. Acta Berl.* **27**, 9–23.
Roothaan, C. C. J. (1951). *Rev. Mod. Phys.* **23**, 69–89.
Stewart, R. F. (1968). *J. Chem. Phys.* **48**, 4882–4889.
Stewart, R. F. (1969). *J. Chem. Phys.* **51**, 4569–4577.
Stewart, R. F. (1971). *Electron Population Analysis with Generalized X-Ray Scattering Factors* (unpublished).
Stewart, R. F. (1972). *J. Chem. Phys.* **57**, 1664–1668.
Stewart, R. F. (1973*a*). *J. Chem. Phys.* **58**, 1668–1676.
Stewart, R. F. (1973*b*). *Acta Cryst.* A**29**, 602–605.

Acta Cryst. (1974). B**30**, 67

On the Application of Phase Relationships to Complex Structures. VI. Automatic Interpretation of Electron-Density Maps for Organic Structures

By M. H. J. Koch

Laboratoire de Chimie Physique et de Cristallographie, Université de Louvain, Schapenstraat 39,
B-3000 *Leuven, Belgium*

(*Received* 3 *August* 1973; *accepted* 6 *August* 1973)

This paper describes a procedure for the automatic interpretation of electron-density maps of organic structures which makes use of the restrictions on the values of the intramolecular bond distances and angles. It is also shown how some chemical-data processing algorithms can be used as an efficient means of introducing chemical information in the process of structure determination.

The last step of a crystal structure determination by direct methods which requires human intervention is the interpretation of E maps. Though graphical display procedures make this easier it is still a time-consuming task, especially when no easily recognizable fragment can be found. We wish to report here some experiments of automatic interpretation of organic structures and compare the results with those obtained by visual interpretation. Organic molecules are a favourable case since one can make full use of the restrictions that exist on the values of the bond distances and angles.

The tests have been carried out starting from the output of the peak-search program of the *MULTAN* system (Declercq, Germain, Main & Woolfson, 1973). This output consists of a list of coordinates of the N highest peaks in the asymmetric unit of an E map. The value of N is usually taken to be 1·2 times the number of atoms in the asymmetric unit. The coordinates in this list are such that the peaks form the largest possible fragments. The interpretation proceeds as follows:

(*a*) A connectivity matrix is set up. This is a symmetrical $N \times N$ matrix where the ijth element is 1 if the atoms i and j are bonded and 0 otherwise. Two peaks are assumed to be bonded if the distance between them lies within the limits D_1 and D_2 (*e.g.* $D_1 = 1{\cdot}1$ Å and $D_2 = 1{\cdot}9$ Å).

Whenever a distance occurs which is less than D_1 the smaller of the two peaks is removed from the list.

(*b*) All angles between adjacent bonds are computed. The smallest of the three peaks is eliminated when the value of the angle lies outside the range A_1–A_2 (*e.g.* 85–145°). An exception is allowed to this rule when a peak which has a connectivity of one is involved in a bad angle. This peak is removed regardless of its weight to ensure that one keeps the largest possible fragments.

(*c*) The number of separate fragments as well as the fragment to which each peak belongs is determined.

(*d*) Symmetry-related positions are tested for all peaks. If a symmetry operation brings peaks from different fragments within bonding distance the two fragments are merged. If this occurs for non-bonded peaks which already belong to the same fragment the smaller peak is eliminated. It is then necessary to repeat step (*c*).

(*e*) Steps (*a*) and (*b*) are repeated.

(*f*) If two fragments were merged, steps (*d*) and (*e*) are repeated until no more changes occur.

(*g*) All peaks which have a connectivity of zero or more than four are removed from the list.

(*h*) Step (*c*) is repeated.

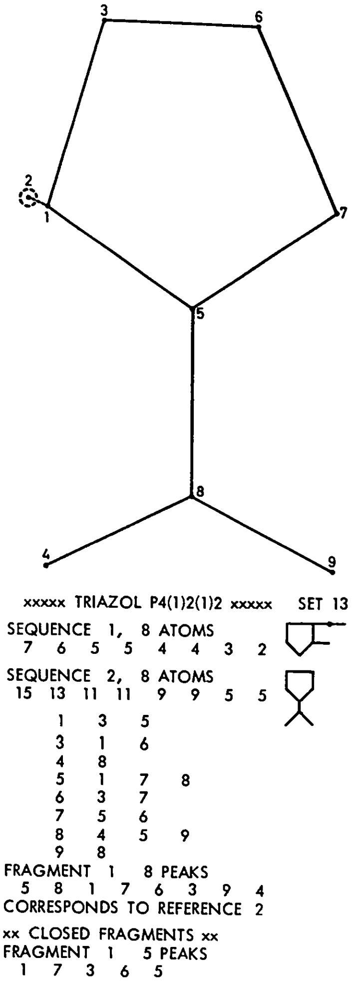
(*i*) The largest fragment or a combination of all fragments which contain five or more peaks is selected as a partial structure which can be input to a phase refinement program.

This procedure can of course be repeated every time a new set of phases is generated from a partial structure by the tangent formula.

Experience with E maps that do not correspond to the correct solution shows that very often no significant fragment can be found. If a meaningful but incorrectly located fragment is present in the map it will be found by the program just as well as by visual interpretation.

The great advantage of visual interpretation comes however from the fact the one can make use of the knowledge of the topology of the molecule. One is usually looking for certain specific features (*e.g.* rings) independently of their exact conformation. This knowledge is also useful in automatic interpretation procedures.

Various algorithms which perform topological searches have been described in connexion with chemical-data processing problems (Lynch, Harrison, Town & Ash, 1971). However as far as we know they have never been applied to crystal structure determination. These algorithms rely mostly on the use of the canonical descriptions of the graph corresponding to the chemical structure. This description can easily be found from the connectivity matrix as follows. The connectivity of each peak is replaced in a cyclic process by the sum of the connectivities of the peaks to which it is bonded until the number of different connectivity

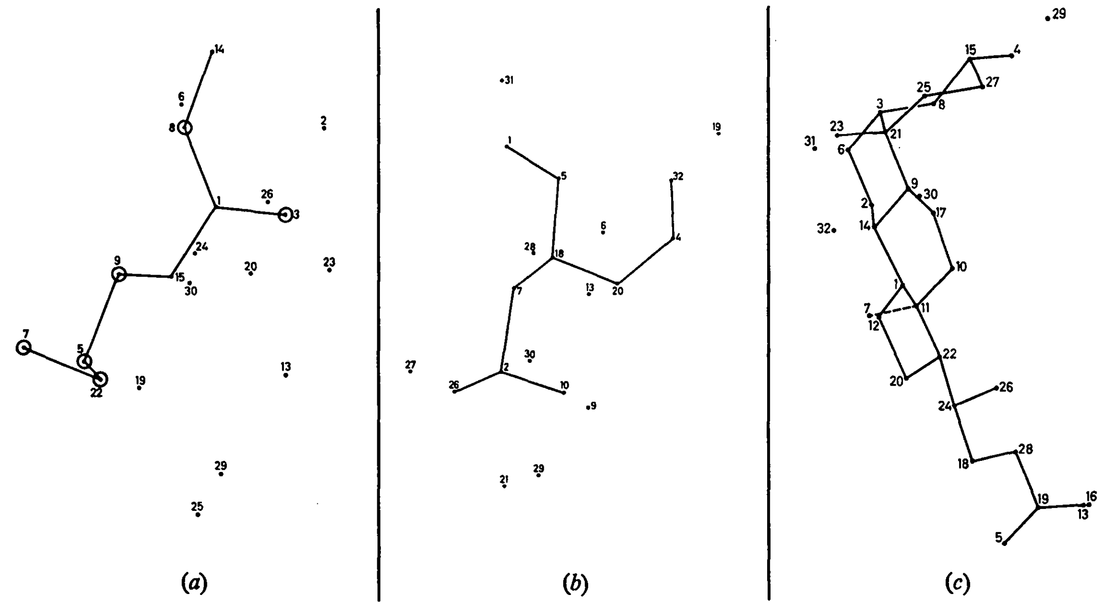


```
xxxxx TRIAZOL P4(1)2(1)2 xxxxx    SET 13
SEQUENCE  1,  8 ATOMS
  7   6   5   5   4   4   3   2
SEQUENCE  2,  8 ATOMS
 15  13  11  11   9   9   5   5
        1   3   5
        3   1   6
        4   8
        5   1   7   8
        6   3   7
        7   5   6
        8   4   5   9
        9   8
FRAGMENT  1    8 PEAKS
  5   8   1   7   6   3   9   4
CORRESPONDS TO REFERENCE  2
xx CLOSED FRAGMENTS xx
FRAGMENT  1    5 PEAKS
 1   7   3   6   5
```

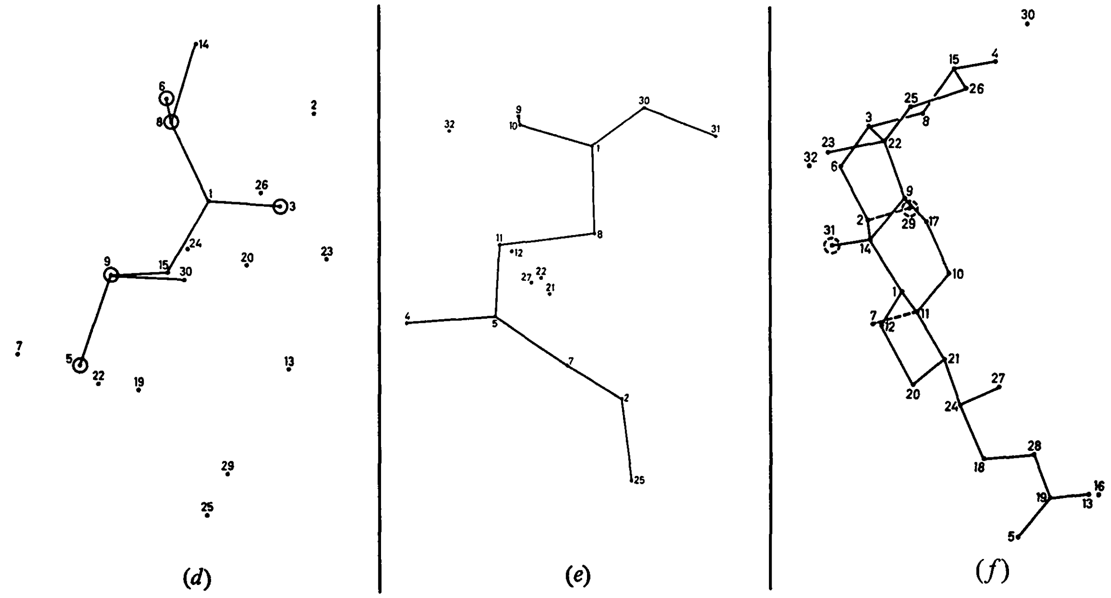
Fig. 1. Computer output for triazole. The sequences are the canonical descriptions of the two possible structures. They are followed by a bonding table and by the results of the interpretation.

values is less than or equal to the number found in the previous cycle. The canonical description is then given by the sequence of connectivities in the last but one cycle.

At the present stage one cannot expect a program to carry out a complete substructure search except to remove the last few spurious peaks in the final map. This is due to the fact that the problem is quite different from a substructure search in a chemical data file. In this case one wishes to find those graphs in the file containing a subgraph which is isomorphic to a given query graph. Owing to the presence of spurious



VISUAL INTERPRETATION



AUTOMATIC INTERPRETATION

Fig. 2. Lithocholic acid: (*a*), (*b*) and (*c*) show the maps at the successive stages of the visual interpretation, and (*d*), (*e*) and (*f*) show the corresponding steps of the automatic interpretation. In the first maps which are of course identical [(*a*) and (*d*)] only the circled peaks are correct. In each case the connected fragment has been used as a partial structure to obtain the next map.

peaks the question that arises in the interpretation of E maps is to discover whether the fragment found in the map contains any subgraph which is isomorphic to any subgraph in the structure. Another difficulty is that one cannot make use of the knowledge of the type of atoms and bonds. This is a major drawback if one wishes to use the existing substructure search algorithms (*e.g.* Sussenguth, 1965). However to simplify the problem one can make use of the 'closed fragments' which exist in many organic structures, *i.e.* fragments in which all peaks have at least a connectivity of two. The closed fragment corresponding to a given structure can be found by removing iteratively the peaks that have a connectivity of one until a stable structure is reached. In the present version of the program any fragment can be encoded to form a query graph. It is first compared with all the fragments which were found and with all the closed fragments which can be derived from these.

This kind of information is at present only used to give an immediate indication of the main features of the map and to make it easier to decide which map is eventually worth a more detailed visual interpretation. Experiments are, however, being carried out to determine the best way of using this information in order to computerize the whole decision process.

A few cases will be described here and the results compared with those obtained by visual interpretation.

Triazole ($P4_12_12$; $C_2H_2O_2N_4$; 1 *molecule in the asymmetric unit*)

The problem in this very simple case was to decide between two possible structures. The descriptions of both isomers were input as query fragments. Fig. 1 shows the output of the most probable E map. The spurious peak (2) was removed because the distance between peak 1 and 2 is too short (0·15 Å). The limits set in the program were: $D1=1{\cdot}1$ Å; $D2=1{\cdot}95$ Å; $A1=85°$; $A2=145°$.

3,3-*Dimethyl*-4,5,9,10,11,12-*hexacarboxymethyl-tetracyclo*[7.2.1.0^{2,4}. 0^{2,8}]*dodeca*-5,7,10-*triene* ($Pn2_1a$; $C_{26}H_{28}O_{12}$; 1 *molecule in the asymmetric unit*)

32 maps were computed and none showed a recognizable fragment. After this first unsuccessful attempt the most probable sets were carefully analyzed.

A fragment of eight peaks, one of which turned out to be spurious, was found in the second most probable set. Several cycles of tangent refinement gave the complete structure.

With the automatic interpretation a fragment of five correct peaks was selected. After a number of tangent-refinement cycles during which increasingly large partial structures were picked out automatically the complete structure was obtained.

Lithocholic acid ($P2_12_12_1$; $C_{24}H_{40}O_3$; 1 *molecule in the asymmetric unit*)

Again, no straightforward solution was found for this structure. Fig. 2 shows the various stages of visual and automatic interpretation. It should be noticed that in the present stage of development of the program there are several spurious peaks which are left at the end of the process. They can however be removed by any substructure search technique since the correct structure is a subgraph of the fragment which is found.

Conclusion

Our results show that the constraints which exist on bond distances and angles are sufficient to make the determination of organic structures fully automatic.

Experience with more than fifteen structures shows that there were no structures which could be solved automatically and not visually but conversely no structure which could be solved visually failed to be solved automatically. The time required to interpret maps with up to 99 peaks is approximately five seconds on an IBM 370/155.

I am indebted to Professor M. M. Woolfson and Dr P. Main for their encouragement and suggestions for further developments of this method. I thank Drs G. Germain, G. Evrard and H. Henke for providing the data for these tests, Professor M. Van Meerssche for his interest in this work, and the Fonds National de la Recherche Scientifique for a fellowship. All calculations were carried out at the Computing Centre of the University.

References

Declercq, J. P., Germain, G., Main, P. & Woolfson, M. M. (1973). *Acta Cryst.* **A29**, 231–235.

Lynch, M. F., Harrison, J. M., Town, W. G. & Ash, J. E. (1971). *Computer Handling of Chemical Structure Information.* London: MacDonald; New York: American Elsevier.

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