have been made for distances and angles involving these atoms. The coordinates of the hydrogen atoms are listed in Table 5.

A stereoscopic view (Johnson, 1965) showing the packing of the molecules is given in Fig. 2. There are no short contacts between molecules. The shortest major-atom-to-hydrogen intermolecular distance is 2·30 Å between H(52) of the base molecule and O(25) of the molecule in translational position $(x,y,z-1\cdot0)$. The shortest hydrogen-to-hydrogen intermolecular distance is 2·38 Å between H(52) in the base molecule and H(49) of the molecule in translational position $(x,y,z-1\cdot0)$.

Chemistry

The chemical results of this analysis are illustrated in Fig. 3. It can be seen that the intermediate examined in this analysis would not be suitable for conversion to the desired steroid because of incorrect fusion of the *C–D* ring (*cis* instead of the desired *trans* fusion). The results of the X-ray analysis are in direct opposition to the result predicted by analogies present in the organic literature. The key step in preparing the *C–D* ring stereochemistry involved a stereospecific hydrogenation. Strong and relatively direct analogies in the organic literature indicated that such a hydrogen would undoubtedly result in a *trans* ring fusion. The undisputable evidence that these analogies were incorrect again amplifies the desirability of the use of X-ray analysis in organic chemistry.

References

DUCHAMP, D. J. (1964). *Amer. Cryst. Assoc. Meeting*, Bozeman, Montana. Paper B-14, p. 29.

HOWELLS, E. R., PHILLIPS, D. C. & ROGERS, D. (1950). *Acta Cryst.* **3**, 210.

International Tables for X-ray Crystallography (1962). Vol. III, p. 202–207. Birmingham: Kynoch Press.

JOHNSON, C. K. (1965). *ORTEP*. ORNL-3794, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

WILSON, A. J. C. (1942). *Nature, Lond.* **150**, 152.

Acta Cryst. (1970). B**26**, 2112

The Use of Integer Programming to Solve Crystal Structures

BY R. J. DAKIN

Mathematics Department, University of Papua and New Guinea, P. O. Box 1144, *Boroko, Territory of Papua and New Guinea*

(*Received* 15 *December* 1969)

This paper describes two classes of integer programming formulations of the phase problem for centrosymmetric crystals projected on to one dimension; they provide a much more complete formulation of the basic problem than do earlier integer programming approaches. One class of formulation seeks to match the structure factor expressions in terms of the phase variables with expressions in terms of atom positions; the second class matches electron-density expressions rather than structure factors. The basic advantage of this approach over more traditional methods is that it enables us to find a global minimum of the discrepancy function rather than a local minimum, obviating the need to work from many initial solutions in turn. Experience so far is limited to small artificial examples which have been solved successfully in all cases. Computational difficulties seem likely to limit the size of structure which can be solved in this way. The extension to three-dimensional structures with or without centrosymmetry is straightforward, but leads to large integer programming problems whose solution is probably beyond the scope of currently available computers and integer programming algorithms.

Integer programming problems are linear programming problems (*i.e.* the minimization of a linear function of non-negative variables subject to linear constraints) in which some of the variables are required to be integers. There are several published algorithms for the solution of such problems, though the solution of large problems is still in doubt.

The discrete nature of the phases of structure factors in centrosymmetric crystals suggests the use of integer programming in the solution of these structures. Freeman, Sime, Bennett, Dakin & Green (1963) showed that integer programming formulations could be used to express such conditions as that phases should be π or $-\pi$, electron density should be non-negative and that it should possess a specified number of peaks; Dakin (1966*a*) gives a corrected and expanded version of this approach. It appears, however, that non-negativity is not a sufficient criterion and that formulations which give adequate recognition of peaks lead to large problems which are difficult to solve. The present approach is a much more direct and complete representation of the problem. Integer variables are introduced

to represent the signs of structure factors and the positions of atoms. The integer programming problem becomes that of finding values of sign and position variables which match as closely as possible the effects of the measured structure factors with the effect of the atoms known to be present. Two classes of formulation have been developed. The first seeks to match structure factors and is therefore related to the least-squares method, the major differences being that a linear criterion of fit rather than the squared discrepancy must be used, and that the possibility of converging to a local but non-global minimum is avoided. Both formulations were developed in 1964 but have previously appeared in unpublished reports only (see Dakin 1964, 1966*b*). Both formulations have been successfully applied to the solution of artificial test examples involving four or five atoms, but lack of access to large scale computing facilities since then has prevented further testing.

The paper is mainly concerned with the one-dimensional centrosymmetric case. The extension to two- or three-dimensional and non-symmetric structures is straightforward but leads to very large problems; it remains to be seen whether the solution of such problems will prove computationally practicable.

Formulation in terms of structure factors

First we need to express a structure factor F_{h00} (hereafter written as F_h) in terms of its sign and modulus. The sign is represented by an integer variable s_h which can take one of the two values 0 (representing minus) or 1 (representing plus). In integer programming terms we have:

$$F_h=(2s_h-1)E_h\,, \tag{1}$$

$$0\le s_h\le 1 \tag{2}$$

and

$$s_h \text{ integral.} \tag{3}$$

E_h in (1) is a constant which, in practice, will be an approximate value for F_h derived from X-ray diffraction data. Hence (1) expresses F_h as a linear function of the variables s_h.

Secondly we need to express the structure factor F_h as a linear function of atom position variables. We have

$$F_h=\sum_j 2f_{jh}\cos(2\pi hx_j) \tag{4}$$

where x_j is the position of the jth atom in the unit cell and f_{jh} is the (h00) scattering factor for the jth atom. This is a highly non-linear function of the positions x_j but it can be made linear by the use of integer variables to represent the positions of the atoms. A simple way of doing this is to approximate the value at x by the value at the nearest of a set of grid points denoted by x_i and define integer variables p_{ri} to represent the number of atoms of type r near position x_i. By making the grid sufficiently fine the approximating error is not important. Equation (4) then becomes

$$F_h=\sum_r\sum_i 2f_{rh}\cos(2\pi hx_i)p_{ri} \tag{5}$$

where

$$\sum_i p_{ri}=n_r\,, \tag{6}$$

$$p_{ri}\ge 0 \tag{7}$$

and

$$p_{ri} \text{ integral} \tag{8}$$

where there are n_r atoms of type r in the half unit cell.

Our problem then becomes one of finding values of s_h and p_{ri} which satisfy equations (2), (3), (6) and (7) and minimize the discrepancies between the corresponding expressions (1) and (5). A least-squares criterion cannot be expressed in terms of linear inequalities but there are several criteria which can be so expressed. One such criterion is to minimize a weighted sum of the moduli of the discrepancies.

If, for each h included in the formulation, we introduce error variables e_{1h} and e_{2h} which satisfy

$$e_{1h}\ge 0,\ e_{2h}\ge 0 \tag{9}$$

and

$$(2s_h-1)E_h-\sum_r\sum_i 2f_{rh}\cos(2\pi hx_i)p_{ri}=e_{1h}-e_{2h} \tag{10}$$

then the smallest value which $(e_{1h}+e_{2h})$ can take is

$$\left|(2s_h-1)E_h-\sum_r\sum_i 2f_{rh}\cos(2\pi hx_i)p_{ri}\right|$$

i.e. the modulus of the discrepancy between the two expressions (1) and (5) for F_h.

To complete the formulation we introduce positive weights w_h and note that one of the signs (s_g, say) can be selected arbitrarily (minus, say). We shall suppose that there are t different types of atom present, n grid points and that, apart from g, a subset S of the structure factors is considered. The complete formulation is then as follows.

Find values of s_h $(h\in S)$, p_{ri} $(r=1,\ldots,t;\ i=1,\ldots,n)$, e_{h1} and e_{2h} $(h\in S,\ h=g)$ which minimize

$$z=\sum_{h=g,h\in S} w_h(e_{1h}+e_{h2}) \tag{11}$$

and satisfy

$$(2s_h-1)E_h-\sum_{r=1}^{t}\sum_{i=1}^{n}2f_{rh}\cos(2\pi hx_i)p_{ri}=e_{1h}-e_{2h}(h\in S) \tag{10}$$

$$-E-\sum_{r=1}^{t}\sum_{i=1}^{n}2f_{rg}\cos(2\pi gx_i)p_{ri}=e_{1g}-e_{2g} \tag{10'}$$

$$\sum_{i=1}^{n}p_{ri}=n_r\ (r=1,\ldots,t) \tag{12}$$

$$0\le s_h\le 1\ (h\in S) \tag{2}$$

$$s_h \text{ integral } (h\in S) \tag{3}$$

$$p_{ri} \geq 0 \ (r=1, \ldots, t;\ i=1, \ldots, n) \qquad (7)$$

$$p_{ri} \text{ integral } (r=1, \ldots, t;\ i=1, \ldots, n) \qquad (8)$$

If there are m members of S then this formulation contains $(3m+2+nt)$ variables and, apart from the individual bounds (2), $(m+t+1)$ constraints. It can readily be rearranged to include in the initial basis, e_{1h} $(h=g, h \in S)$ and one p_{ri} for each r to give $[2m+1+t(n-1)]$ non-basic variables.

The actual size of integer programming problem required to solve a given structure will depend on the number of structure factors included in the formulation and the number of grid points; this raises the question of how many of each should be included for a given structure. The computational difficulties in solving the integer programming problem are such that one should not expect to obtain a refined structure immediately; it will probably be more efficient to use the coarsest formulation which will allow us to correctly resolve the structure and refine this as a subsequent step.

There seems to be no firm basis which would allow one to determine theoretically the coarsest model which will work, but some rough rules are apparent. Presumably one would need to include at least as many structure factors as there are atoms whose positions are to be determined – somewhat more than this is probably necessary. For one of the test examples twice the number of atoms appeared to be barely sufficient. If one allows for six grid points per cycle of the highest order structure factor (h_1, say) this gives $3h_1$ grid points for the half unit cell. Hence for a cell with no symmetry (additional to centrosymmetry) with $M(=\sum n_r)$ atoms in the half cell it might be reasonable to include the first 1·5 M structure factors and 4·5 M grid points, giving a problem with $[M(3+4{\cdot}5t)-t-1]$ variables and $1{\cdot}5M+t$ constraints. This is a tentative suggestion which requires testing. Probems with additional symmetry will require the inclusion of fewer structure factors for a given total number of atoms, but high order components necessitating a fine grid will still be included, since some low order components will be absent because of symmetry.

Table 1 shows the size of integer programming problems arising from problems with no additional symmetry, based on the above assumptions. The Table gives data storage requirements for the particular program used to obtain the results below. If internal storage only is used then storage could become a problem for more than about 40 atoms in the half unit cell if there is only one atom type, and somewhat fewer atoms if there are more types. With current trends in computer development, storage should become progressively less of a problem. It may well be that one can correctly resolve the structure using a compressed model in which some types are amalgamated, with a compromise set of scattering factors.

At the present state of the art computation time is likely to be the limiting factor. On the basis of known experience one would not be very confident about solving problems with more than about 20 (1 type) or 7 (4 types) atoms in the half unit cell. However, greater problems might be worth attempting, especially since some integer programming algorithms (such as that described by Dakin, 1965) will reach a solution, not necessarily optimal, quite quickly. It would be worth investigating whether these initial solutions correctly resolve the structure in a high proportion of cases. For the small artificial examples so far investigated initial solutions were substantially correct.

Formulation in terms of electron density

Our approach so far has been to express the structure factors in terms of both sign and atom position variables and to minimize the discrepancy. An alternative approach is to obtain two expressions for electron density – one in terms of sign variables and measured structure factors and the other in terms of atom positions and to minimize some measure of the total discrepancy between the two expressions over some set of points ξ_j; $j=1, 2, \ldots, N$ not necessarily the same points x_i which are used for atom positions). As before one may either minimize the sum of moduli of discrepancies or the largest discrepancy. The expression in terms of sign variables is obtained by replacing $F_h \cos(2\pi h\xi_j)$ by $(2s_h-1)E_h \cos(2\pi h\xi_j)$ in the usual expression for electron density. In terms of position variables the electron density at ξ_j is given by $\sum_{r=1}^{t}\sum_{i=1}^{n} A_r(\xi_j-x_i)p_{ri}$, where the function $A_r(\xi-x)$ gives the contribution to the electron density at ξ of an atom of type r centred at x.

One cannot predict which approach is likely to prove more successful in practice: both should be tried. If the atom position grid and the number of Fourier terms are the same and density is sampled at the same number of points as there are Fourier coefficients included then both formulations give the same size of problem. Just how the smallest practicable formulations for the two approaches would compare remains to be seen. The

Table 1. *Computer storage requirements*

	1 Atom type			4 Atom types		
Atoms	Constraints	Variables	Storage	Constraints	Variables	Storage
4	7	38	300	10	79	1000
10	16	73	1400	19	205	4300
20	31	148	5000	34	415	15000
40	61	298	19000	64	835	55000
100	151	748	115000	154	2095	330000

first approach has the advantage that it makes no assumptions about structure factors which are omitted; the second approach assumes, in effect, that they are of zero magnitude. The ability to put a weighting on different structure factor errors may also be an advantage for the first approach where one has reasons for believing that some measurements are more accurate than others.

A further variant is to use piecewise linear approximations rather than step approximations to the functions $E_h \cos(2\pi hx)$ (first approach) or $A_r(\xi_j - x)$ (second approach). The integer position variables p_{ri}, indicating the presence of an atom in a cell of the grid, are augmented by continuous position variables q_{ri}, indicating the position of the atom within the cell; q_{ri} must satisfy the constraints

$$0 \le q_{ri} \le p_{ri} \ (r = 1, \ldots, t;\ i = 1, 2, \ldots, n) .$$

If we replace $\cos(2\pi hx_i)p_{ri}$ in (10) by $a_{hi}p_{ri} + b_{hi}q_{ri}$ then $\cos(2\pi hx)$ can be represented in cell i by an arbitrary straight line segment whose position and slope are determined by the constants a_{hi} and b_{hi} respectively. These constants might be chosen to minimize the maximum approximating error. For a given atom position grid size this gives a smaller approximation error, but requires more constraints and variables. One can take advantage of this and use a coarser grid, but it appears that for a given level of approximating error the step approximation will still lead to a smaller integer programming problem.

Extension to three-dimensional and asymmetric problems

Both formulations and their variants can be extended to two- or three-dimensional structures which are not necessarily centrosymmetric. The resulting problems are very large; their solution may not be practicable by currently available techniques and computers – although this point is not definitely established. Developments in solution algorithms and large scale computers will improve this situation.

The three-dimensional centrosymmetric case involves very little extension other than the use of a three-dimensional rather than a one-dimensional grid. We shall also need to replace terms of the form $\cos 2\pi hx$ by terms of the form $\cos 2\pi(hx + ky + lz)$. The size of problem, however, grows much more rapidly with the number of atoms than the one-dimensional model. For a given number of atoms one will presumably require about three times as many structure factors, as there are three times as many coordinates to be determined; for a given grid spacing the number of points will be nearly cubed. The use of piece-wise linear approximations should have a decided advantage in three dimensions, since the penalties in obtaining a finer grid are so great.

Non-centrosymmetric problems with continuous phase variation will give rise to sinusoidal terms which can be handled by step or piecewise linear approximations as before. If necessary it is possible to introduce constraints on the spatial relationship of atoms in three-dimensional formulations. For example, one could specify a spacing of at least D by putting an upper bound of one on the number of atoms in spheres of diameter D.

Experience

In order to test the methods some small artificial examples were constructed. As far as I know no genuine crystallographic problems have been tackled by these methods.

Peaks of the form

$$A(\xi - x) = B \exp[-k(\xi - x)^2]$$

were used, where B and k are constants. Peaks were placed randomly and structure factors were calculated and then rounded off, to give something of the effect of experimental errors in a real example. Errors introduced in this way are rectangularly distributed with a maximum error of approximately 0·001 times the zero order term. Problem 4 is an exception, as we shall see.

Table 2. *Test examples*

	Atom parameters				Atom positions				
	Type 1		Type 2		Type 1			Type 2	
No.	B	k	B	k	a	b	c	a	b
1–3	10	790	7	632	0·111	0·302	—	0·207	0·414
4	10	790	7	632	0·083	0·400	—	0·200	0·300
5	10	800	7	620	0·131	0·343	0·496	0·321	0·364

Table 3. *Formulation details*

		Structure factors			Grids		Problem size	
No.	Formulation	Total	Fixed sign	Atom types	n	N	Constraints	Variables
1	Electron density	0–17, 19, 20	0·7	2	31	15	17	93
2	Electron density	0–17, 19, 20	0·7	1	51	21	22	89
3	Structure factors	1—8	1·7	2	31	—	10	74
4	Electron density	0–23	0·9	2	31	21	23	103
5	Structure factors	1–10	1	2	41	—	12	99

Table 4. *Result summary*

No.	Sign errors	Atom position errors 1*a*	1*b*	1*c*	2*a*	2*b*	Grid spacing	Interations	Computation time for KDF9 (sec)
1	1, 19, 20	0·011	0·002	—	0·007	0·003	0·017	460	40
2	17	0·001	0·002	—	0·007	0·004	0·010	920	130
3	8	0·011	0·035	—	0·040	0·014	0·017	3510	220
4	–	0·000	0·000	—	0·000	0·000	0·017	2240	290
5	–	0·006	0·006	0·004	0·016	0·014	0·013	440	40

Details of the problems and formulations are given in Tables 2 and 3. Nos. 1-3 are three different formulations of the same problem. No. 1 is a straightforward application of electron-density matching; it includes a large number of Fourier terms for the number of atoms and the grid spacing is rather coarse for the higher terms; this may explain why s_{19} and s_{20} were incorrectly determined. The formulation for No. 2 uses only one peak type (with parameters $B=8{\cdot}5$, $k=711$) to reduce the problem size. No. 3 includes a comparatively small set of signs with an appropriate grid spacing. No. 4 was constructed so that the atoms fell on grid points to see whether exact results could be obtained for such a case. Computed structure factors were not rounded off but included to six figure accuracy in this case.

The results are given in Table 4. A tree search integer programming algorithm (Dakin, 1965) was used. The solution algorithm failed to terminate in each case, so the solutions reached may not be optimal, but the significant feature of the results is that structures were correctly resolved in all cases except, possibly, No. 3 which is the only case in which the atom position errors exceed the grid spacing. Even here the resolved structure may well be near enough to reduce to the correct structure when refined. It would appear that this formulation included barely enough structure factors; a slightly altered version of the integer programming algorithm failed to find a solution within 10 minutes. In all cases the terms whose signs were incorrect were comparatively small.

The amount of computation is indicated in the last two columns, which give the number of iterations of the Simplex method and the central processor computation time for the KDF9. In the absence of other experience no great significance can be placed on the differences in the computations required for the different formulations. The computation requirements are likely to grow quite rapidly with increasing size of problem; it would be well worth finding out just how rapidly.

Conclusions

The methods we have described give promise of providing automatic methods for solving small centrosymmetric problems and could possibly be useful for more general problems. Just how reliable the method is and how far it can be pushed using modern large-scale computers remains to be seen. It would be well worth following up.

I am indebted to Dr H. C. Freeman and Dr J. Sime who introduced me to the phase problem and to Professor J. M. Bennett who first suggested that integer programming might be applicable. This research was supported, in part, by Air Force Office of Scientific Research Grants AF-AFOSR-62-402 and AFOSR-64-686, and was carried out at Sydney University.

References

DAKIN, R. J. (1964). *Basser Computing Department Technical Report* No. 31, Sydney Univ.

DAKIN, R. J. (1965). *Computer J.* **8**, No. 3, 250.

DAKIN, R. J. (1966*a*). Ph. D. Thesis, Ch. 5. Basser Computing Department, Sydney Univ.

DAKIN, R. J. (1966*b*). Ph. D. Thesis, Ch. 6. Basser Computing Department, Sydney Univ.

FREEMAN, H. C., SIME, J. G., BENNETT, J. M., DAKIN, R. & GREEN, D. (1963). *Symposium on Crystallographic Computation Methods*, I.U.Cr., Rome 1963.