Acta Cryst. (1978). A**34**, 782–791

Experience with Fast Fourier Least Squares in the Refinement of the Crystal Structure of Rhombohedral 2-Zinc Insulin at 1·5 Å Resolution

BY N. W. ISAACS* AND R. C. AGARWAL†

IBM Thomas J. Watson Research Center, PO Box 218, *Yorktown Heights, NY* 10598, *USA*

(*Received* 2 *February* 1978; *accepted* 5 *April* 1978)

The newly developed fast Fourier least-squares algorithm [Agarwal (1978). *Acta Cryst.* A**34**, 791–809] has been used to refine the structure of rhombohedral 2-Zn pig insulin at a resolution of 1·5 Å. The CPU time required for each cycle of refinement was about 3 min on an IBM 370/168 computer. After 67 cycles of refinement the conventional R factor was 0·113 for 11 890 terms. The procedure used in the refinement is described and an analysis of the results presented. From the experience learned in this refinement a procedure is suggested which should provide faster convergence and require considerably fewer refinement cycles.

Introduction

In recent years a number of protein crystal structures have been refined with high-resolution ($d \leq 2$ Å) diffraction data. With one exception, the refinement technique used has been either the difference-Fourier method (Freer, Alden, Carter & Kraut, 1975; Moews & Kretsinger, 1975; Adman, Sieker & Jensen, 1975; Chambers & Stroud, 1977) or the real-space refinement method of Diamond (1971, 1974) (Huber *et al.*, 1974; Deisenhofer & Steigemann, 1975; Bode & Schwager, 1975; Takano, 1977*a*,*b*). The exception is the structure of rubredoxin which has been refined by conventional least squares (Watenpaugh, Sieker, Herriot & Jensen, 1973). Although the least-squares method provides a considerably more accurate structure, it is unlikely that a conventional method of least-squares refinement will be used routinely for protein structures, since the computation time required is roughly proportional to NM^2 where N is the number of data and M the number of parameters in the refinement.

A new algorithm for least-squares refinement has been developed (Agarwal, 1978) which is especially applicable to the refinement of structures of quite large molecular weights. This algorithm uses fast Fourier transforms (FFT) at all possible stages of the computation, and the computation time required is approximately proportional to $N \log N$, where N is the number of reflections. In this paper we describe our experience in using the method to refine the crystal structure of rhombohedral 2-Zn pig insulin at 1·5 Å resolution. On an IBM 370/168 computer, this refinement took an average of 3 min CPU time per cycle. In a parallel study, this structure has been refined by difference Fourier methods (Dodson, Dodson, Hodgkin & Vijayan, 1978) and we have used these results to verify the correctness of our own procedure. This comparison revealed some notable differences; we shall describe some of these in relation to the use of the least-squares method. The refinement of the insulin structure is not yet complete and a full description of the refined structure will be presented separately.

Experimental

Rhombohedral 2-Zn insulin crystallizes in the space group $R3$ with cell parameters $a_R = 49{\cdot}0$ Å, $\alpha_R = 114{\cdot}8°$; $a_H = 82{\cdot}5$, $c_H = 34{\cdot}0$ Å. There are two zinc atoms and two insulin molecules each of MW 5780 daltons in the asymmetric unit. The stoichiometric solvent content of the asymmetric unit is equivalent to 280 water molecules. Data to a resolution of 1·5 Å which had been collected on a four-circle diffractometer were supplied by D. C. Hodgkin and G. G. Dodson. Of the 13 668 data in the 1·5 Å sphere, 10 119 had $|F| > 2\sigma(|F|)$, where the standard deviations were estimated from counting statistics. In all our calculations we have used the hexagonal setting of the rhombohedral cell.

The crystal structure of 2-Zn insulin was first reported at a resolution of 2·8 Å (Adams *et al.*, 1969) and later at a resolution of 1·9 Å (Hodgkin, 1974). Coordinates derived from this map have been used as a starting point for the difference Fourier refinement of the structure at 1·5 Å resolution (Dodson *et al.*, 1978). We have used an entirely independent set of coordinates to start the least-squares refinement. The phase extension and refinement method described by Sayre (1972, 1974) was used to obtain phases at 1·5 Å

*Present address: Department of Biochemistry, University of Melbourne, Parkville, Victoria 3052, Australia.

†Present address: Centre for Applied Research in Electronics, Indian Institute of Technology, New Delhi 110029, India.

resolution directly from the 1·9 Å isomorphous phases (Cutfield *et al.*, 1975). From an electron density map computed with these phases, the positions of 743 protein atoms were obtained. These coordinates were modified to produce a structure with regular geometry (Dodson, Isaacs & Rollett, 1976). After five rounds of difference Fourier refinement, the conventional reliability factor $[R = \sum (k|F_o| - |F_c|)/\sum k|F_o|]$ for 13 393 terms in the resolution range 10 Å $\geq d \geq$ 1·5 Å had been reduced from an initial value of 0·385 to 0·251 (unregularized coordinates). The number of atoms placed had increased to 853, including 74 solvent atoms. After regularization, these coordinates were used to start the least-squares refinement.

The refinement was carried out using the fast Fourier least-squares algorithm (Agarwal, 1978). The function minimized is $P = \sum_{hkl} W_{hkl} (|F_o| - k|F_c|)^2_{hkl}$ where W_{hkl} is a weighting function and k is the scale factor. In this algorithm, fast Fourier transforms (Cooley & Tukey, 1965; Ten Eyck, 1973) are used to compute the structure factors and the derivatives of P with respect to the atomic parameters. The normal matrix was approximated as a nearly diagonal matrix [the only off-diagonal terms considered were between non-orthogonal coordinates (x and y in the space group R_3) of the same atom]. Each cycle was a refinement of either the atomic coordinates or the individual isotropic temperature factors. The shifts produced by the least-squares procedure were multiplied by a step size (or scalar) and this step size was also refined to give the optimum convergence rate (Agarwal, 1978). The CPU time and core-storage requirement varied depending on the grid size, which is dependent on the resolution of the data used in the refinement. In the latter stages of the refinement, each cycle required approximately 3 min CPU time and 600 K words of data and program storage on an IBM 370/168. Since we had a large virtual memory (directly addressable storage) available on this machine we wrote the program assuming that all the arrays were residing in the main core. The program has since been modified for other computer systems to use random-access disk files for storage of the large arrays. It has operated successfully on a DEC-10 computer at the University of York where it required 90 K words of core storage and 12 min of run time (CPU time plus paging and other operating costs) for each cycle.

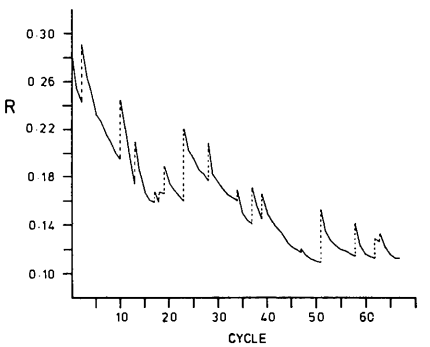

Fig. 1. Plot of R value *vs* cycle number.

The course of the refinement is shown in Fig. 1, which plots the decrease in the R factor with cycle number. The refinement gave consistent convergence from an initial R factor of 0·282 for 6572 terms at 1·83 Å resolution to a final R of 0·113 for 11 890 terms at 1·5 Å resolution (0·148 for all 13 424 terms in the resolution range 14 Å $\geq d \geq$ 1·5 Å). The breaks in the curve of Fig. 1, where the R factor increases, were caused either by a modification of the structure (such as regularization of the geometry or addition/deletion of atoms) or by a change in the number of data used in the refinement. Tables 1 and 2 summarize these details. The final refined model consisted of 813 protein atoms (including 2 zinc atoms and 10 atoms assigned half occupancy) and 264 solvent atoms (of which 82 were assigned half occupancy). Of the 67 cycles of refinement, 43 were on coordinates and 24 on temperature factors. These may seem to be excessively large numbers (even though the total CPU time involved is less than 4 h), but we consciously calculated a large number of refinement cycles under varying conditions in order to learn the optimal procedure (see below).

Limits of data used

As shown in Table 1, varying amounts of data were used, particularly in the initial stages of the refinement. The low-angle data ($d \geq$ 10 Å) were excluded from all the calculations in the first 47 cycles, as these data are most severely affected by the solvent continuum in the crystal. From cycle 48 on, when we felt more confident of our description of the solvent, we included data with $d \leq$ 14 Å.

The variation in the upper limit of $(4 \sin^2 \theta)/\lambda^2$ in the first 19 cycles was experimental as we wished to assess the effect of using a resolution-limited data set on the rate of convergence of the refinement. Since an unrefined set of coordinates may cause relatively large errors in the calculated structure factors of high-resolution terms, we expected that the exclusion of these terms would speed up the convergence. This was borne out in practice. Furthermore, since the grid spacing used in the FFT depends on the resolution of the data (we used an interval of approximately one third of the resolution of the data), this procedure has the added advantage of allowing a coarser grid with a resultant saving in computer time.

As well as limiting the data according to the resolution, we did not include any low-magnitude structure factors in the first 23 cycles. We divided the data into ranges of $\sin \theta/\lambda$, and for each range used only those data with magnitude greater than 10% of the maximum

Table 1. *Summary of the major steps in the refinement*

Cycles	Type	R	Limits $\frac{4 \sin^2\theta}{\lambda^2}$	Number of data	Number of atoms	Regularized (after last cycle)	Revised
1–2	2*XYZ*	0·243	0·01–0·30	6572	853	Yes	
3–10	4*XYZ*,4*B*	0·195	0·01–0·45	11373		Yes	Yes
11–13	2*XYZ*,1*B*	0·174	0·01–0·35	8307	958	Yes	
14–17	2*XYZ*,2*B*	0·159					
18	*XYZ*	0·167	0·01–0·40	9902			
19	*XYZ*	0·167	0·01–0·44	11278	957	Yes	
20–23	3*XYZ*,1*B*	0·160	0·01–0·445	11285		Yes	Yes
24–28	3*XYZ*,2*B*	0·177		13393	922	Yes	Yes
29–34	4*XYZ*,2*B*	0·160		13393	984	Yes	
*35–37	2*XYZ*,1*B*	0·141	0·01–0·45	12409		Yes	Yes
38–39	2*XYZ*	0·145		12185	1029	Yes	
40–47	6*XYZ*,2*B*	0·118		12027			Yes
48–51	2*XYZ*,2*B*	0·110		12050	1113	Yes	Yes
52–58	5*XYZ*,2*B*	0·113		11919	1010	Yes	Yes
59–62	2*XYZ*,2*B*	0·113	0·005–0·45	11942	1055	Yes	
63	*B*	0·125		11820		Yes	Yes
64–67	2*XYZ*,2*B*	0·113		11890	1077		

* From cycle 35 on, hydrogen atoms were included in the structure factor calculation, and restricted data were used (see text).

magnitude for that range. In retrospect this did not seem to have any positive (or negative) effect on the refinement. In cycles 24 through 34 all data were used ($d \leq 10$ Å). We felt a need at this stage to discriminate between reliably measured data and unreliable data (unobserved terms). We had $\sigma|F|$ values only for data in the range $1{\cdot}9 \geq d \geq 1{\cdot}5$ Å. Because of this incomplete knowledge of the accuracy of all the data, we could not discriminate against data on the basis of $|F| \geq 2\sigma|F|$. In fact, even if we had this information available we would have been reluctant to use it as it would have removed too many data from the refinement. We tried then to find an empirical method of discriminating against the poorly measured data and chose to select data on the ratio of their observed and calculated structure-factor magnitudes. For the remaining cycles all data within the given resolution limits were used subject to the condition that $1/t \leq |F_o|/|F_c| \leq t$ where $t = 2$ in cycles 35 to 37 and 1·8 otherwise. Our choice of these limits was based on an inspection of a plot of R factor against $2 \sin\theta/\lambda$ (Fig. 2). Normally such a curve should show a continuous increase in R with angle (Luzzati, 1952), whereas we found a sharp increase in the slope of the curve starting at about 1·9 Å resolution. By restricting the data to the condition given, the curve showed a more normal appearance without the loss of too many terms. The effect of the restriction was to remove from the refinement terms which were primarily of high angle and low-magnitude F_o, *i.e.* 'unobserved' terms. But we also removed some low-angle terms where the large discrepancy between the observed and calculated structure factors was perhaps due to our inadequate description of the protein and particularly of the solvent

Table 2. *Numbers of atoms included in structure factor calculations and refinement*

The numbers in brackets give the number of atoms assigned half occupancy.

Cycle	Protein (refined)	Solvent (refined)	Hydrogen (not refined)
0	779	74	–
11	802	156	–
24	781	141	–
29	800	184	–
35			718
38	808 (2)	221	734
48		305 (84)	737
52	809 (2)	201 (18)	719
59		246 (64)	719
64	813 (10)	264 (82)	749

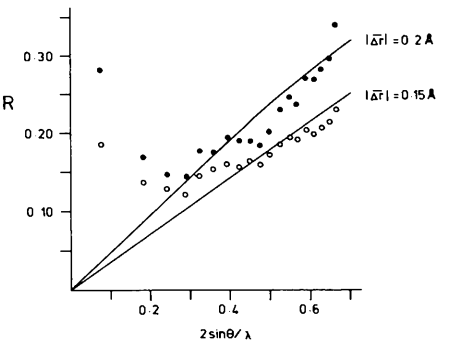

Fig. 2. The distribution of the R value with $2 \sin\theta/\lambda$, calculated after 37 cycles of refinement. The full circles are for all the data and the open circles for data satisfying the criterion $1/t \leq |F_o|/|F_c| \leq t$, where $t = 1{\cdot}8$.

continuum. For this reason we feel that this procedure should be used with care and we would not recommend it as a general practice, although in the present case we have not found any errors in the structure caused by its use.

Scaling and weighting scheme

The effects of the high solvent concentration of protein crystals on low-resolution data ($d > 10$ Å) present difficulty in calculating the scale to apply to the calculated structure factors. This problem is usually overcome by omitting the low-angle data from the calculation or by applying a multiparameter scale factor to obtain the best fit between the distribution of $|F_o|$ and $|F_c|$ over the whole range of sin θ/λ (Moews & Kretsinger, 1975; Chambers & Stroud, 1977). We found that a scale factor estimated by the relation $k = (\sum|F_o|\,|F_c|)/\sum|F_c|^2$ where the sum includes all the data used in the refinement, gave satisfactory results.

Luzzati (1952) has shown that a plot of R factor against sin θ for various mean errors in the coordinates gives a series of curves which are linear for small coordinate errors, and approach an exponential curve for large errors. It is preferable then, when the initial coordinate errors are large, to place more weight in the refinement on the low-angle terms rather than the high-angle terms where the errors between the observed and calculated structure factors are more random. We used a weighting scheme of the form $w_{hkl} = (2 \sin \theta/\lambda)^p$. For cycles 1 through 7, p was given a value of $-1{\cdot}0$, then for cycles 8 through 12 relaxed to $-0{\cdot}5$ and finally 0·0 (*i.e.* unit weights) for the remainder of the refinement.

Regularization of the structure

At all stages it was necessary to apply a regularization procedure to correct the geometry of the structure. At the beginning of the refinement in particular, the shifts applied by the least-squares procedure produced some quite unacceptable stereochemistry in the protein. Possible causes of this effect are the limited amount of data used (12 000 observations and ~4000 parameters), missing atoms, incorrectly placed atoms, missing hydrogens and solvent atoms, the large thermal motion of the protein and the use of a diagonal approximation to the normal matrix which did not account for atom–atom interactions. Of all these causes, the large thermal motion has the most serious effect.

The geometry of the structure was corrected with a program written by Dodson *et al.* (1976). In this procedure the atomic coordinates are adjusted to satisfy four types of constraints, *viz* (1) a constraint to the original coordinates, constraints (2) to an ideal bond length and (3) to an ideal bond angle, and (4) a constraint on planar groups of atoms. The relative weights given to each of the first three constraints may be adjusted by assigning an estimated standard deviation (e.s.d.) to the atom coordinates, bond lengths and bond angles. Assigning low e.s.d.'s to the bonds and angles produces a structure with geometry close to ideal, but at the expense of some large movement of atoms. The parameters we used in regularization during the refinement are given in Table 3. We found that applying tight constraints to bonds and giving all atoms the same e.s.d. caused an oscillatory effect for those atoms with low temperature factors which had refined quickly. To overcome this, for regularization after cycle 28, we assigned e.s.d.'s to the atom coordinates

Table 3. *Details of the regularization procedure*

From cycle 28, σ_{coord} was estimated for each atom according to its B value (see text).

Cycle	R increase	R.m.s. bond (Å)*		E.s.d.		R.m.s. shift (Å)	σ_{coord} (Å)
		Before	After	Bond (Å)	Angle (°)		
2	0·049	0·37	0·07	0·015	3	0·30	0·10
10	0·053	0·29	0·07			0·28	0·10
13	0·036	0·23	0·01	0·02	5	0·17	0·08
19	0·023	0·27	0·06			0·19	0·08
23	0·062	0·25	0·06	0·05	5	0·18	0·08
28	0·032	0·25	0·05			0·21	
34	0·006	0·21	0·06	–	–	0·18	
37	0·030	–	–			–	
39	0·021	0·10	0·02	0·025	2·5	0·09	
51	0·043	0·17	0·02			0·10	
58	0·027	0·20	0·02			0·09	
62	0·016	0·16	0·04	0·05	5	0·06	
63	0·007	–	–	–	–	–	
67	0·013	0·09	0·03	0·05	5	0·03	

* Defined as the r.m.s. deviation of sp^3 C–C bonds from the expected value of 1·54 Å.

according to the temperature factor of the atom. We used the empirical equation $\sigma = 0{\cdot}2(B/8\pi^2)^{1/2}$, which gave an e.s.d. of 0·1 Å for an atom with a B of 20 Å^2. This considerably improved the regularization procedure, as an acceptable geometry for the molecule was obtained with a smaller increase in R value.

Difference Fourier maps

Our use of difference Fourier maps was limited to searching for solvent molecules and to corrrecting possible errors in the protein structure as indicated by abnormally high temperature factors. In searching for new structure we found that a combination of both $|F_o| - |F_c|$ and $2|F_o| - |F_c|$ maps was best. The $|F_o| - |F_c|$ map showed solvent structure quite clearly, but it also gave peaks close to other atoms which may be due either to an incorrectly placed atom or to residual density from hydrogen atoms not included in the structure factor calculation. The $2|F_o| - |F_c|$ map was particularly useful in placing the side chains of poorly defined residues. We did not use the difference Fourier map to verify systematically the correctness of all the atom positions, or to make small manual adjustments. This was deliberate as we wished to assess the ability of the program to make these adjustments. In retrospect, we feel that, had we made fuller use of difference Fourier maps, the refinement would have converged much more readily.

Solvent structure

Whenever possible we included solvent molecules in the refinement. The tightly bound water was visible at a very early stage and we found that, if these atoms were omitted, the temperature factors of adjacent atoms tended to increase in order to account for the residual density. We experienced difficulty in placing water molecules which were hydrogen-bonded to carboxylic acid side chains. Often those water molecules appeared in the difference map with greater densities than the side-chain atoms. By placing only the terminal atoms of the side chain into this density, then building the side chain back to the main chain we placed not only the solvent, but also the side-chain atoms incorrectly. However, it often happened that some of the atoms placed were given positions close to correct positions for other atoms (for example C_β and C_δ in Glu) and this situation was particularly troublesome. In refinement, the behaviour of such atoms was recognizable as never being really satisfactory, but at the same time there was a sufficient amount of truth in the structure to obscure the correct assignment in difference Fourier maps.

With regard to the less-well-ordered solvent structure, we would advise some caution in placing atoms on all the peaks of a difference Fourier map. Many of the waters which we had placed by cycle 48 were found to be incorrect on a closer examination of the temperature factors and coordination geometry. Most of the solvent molecules removed at cycle 52 (see Table 2) had high B values (>60 Å^2) and had been assigned half occupancy.

Hydrogen atoms

The hydrogen atoms of the protein molecule were included in the structure factor calculation from cycle 35 onwards. They were placed according to the expected geometry on a regularized protein structure and assigned B values equal to those of the atoms to which they were bonded. Their positions were not refined and new positions were calculated whenever the geometry of the protein was regularized. Including the hydrogen atoms in the structure factor calculation reduced the errors in the agreement between low-angle terms and also reduced the least-squares shifts calculated for the well defined (low temperature factor) protein atoms. In the absence of the hydrogen atom contribution to the structure factor calculation, there had been a tendency for atoms to move towards the residual hydrogen density.

Disordered structure

We were cautious in assigning disordered structure to the protein as we felt that for residues with large temperature factors, an incorrect description of disorder would disguise the appearance of the correct conformation in difference Fourier maps. The problem is essentially one of the interpretation of weak density.

Table 4. *Residues revised after cycle* 28

Residue	Revised after cycle 28	37	51	63	Mean B for side chain
Molecule 1					
A21 Asn		+	+		46
B21 Glu	+	+	+	+	55
B22 Arg		+	+		47
B27 Thr		+			42
B29 Lys	+	+	+		35
B30 Ala	+	+	+		74
Molecule 2					
A5 Gln		+	+	+	55
A13 Leu	+				33
A17 Glu		+			28
A21 Asn	+	+	+	+	45
B3 Asn				+	49
B4 Gln				+	41
B21 Glu		+		+	48
B22 Arg		+	+		50
B29 Lys	+		+	+	47

In our final model we have described disorder for only two residues, B12 Val of molecule 1 and B29 Lys of molecule 2. We failed to interpret the disordered structure for B4 Gln of molecule 1 and B22 Arg of both molecules 1 and 2. Difference Fourier maps calculated with only these residues removed from the structure factor calculation were unclear. One reason for this was that we had placed solvent molecules in the density of one of the possible conformations. This is discussed more fully in the comparison with the difference Fourier refinement.

Corrections to the structure

Table 4 gives a list of the residues manually corrected after cycle 28. There are only 10 residues where the co-ordinates were modified more than once. In each case, the atoms in the side chain have large temperature factors, appeared weakly in the difference maps and refined poorly. *Most of the time and effort spent in the refinement was given to these poorly defined side chains.*

Results and discussions

We do not intend to provide a complete description of the structure of insulin here. A comparison of our results with those arrived at by difference Fourier refinement (Dodson *et al.*, 1978) showed errors in both analyses and the refinement of the corrected structure is still progressing. We hope the following summary of the results of the least-squares refinement will indicate the overall accuracy of our final model.

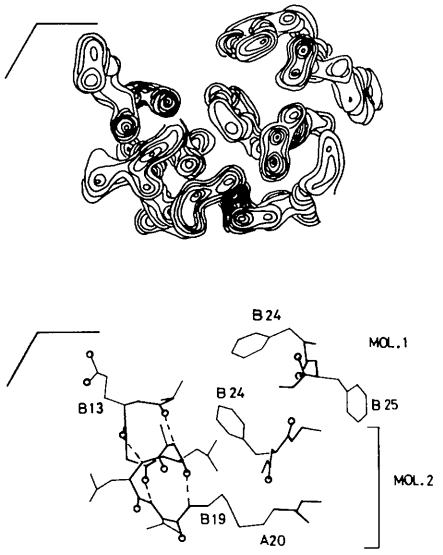

Fig. 3. Superimposed sections of a portion of the final $|F_o|$, α_{calc} electron density map.

Fig. 3 illustrates the quality of the final electron density map. The density corresponding to residues B13–19, B24 and A20 of molecule 2 is shown together with that for residues B24 and B25 of molecule 1. This section of the map was chosen to show the influence of thermal motion on the quality of the map and on the refinement. Residues B13–B19 form part of the α-helix; a well defined structure with little thermal motion in the main-chain atoms ($\bar{B}$ is 9·6 Å^2). The density is clear, unambiguous and approaching atomic resolution. The disulphide bridge (B19–A20, $\bar{B}$ = 12·6 Å^2) is similarly well defined. The side chains of residues B13 Glu and B17 Leu have larger temperature factors ($\bar{B}$ = 23 Å^2 for each) and this is reflected in the map. The aromatic residues B24 Phe of both molecules 1 and 2 are shown at the monomer–monomer contact surface which places constraints on the movement of the side chains. These aromatic rings have low temperature factors ($\bar{B}$ = 10·5 Å^2) and the electron density is extremely clear. By comparison, B25 Phe of molecule 1 is on the surface of the structure and has large thermal motion ($\bar{B}$ = 33·9 Å^2). This was the only residue, of those shown here, which refined poorly. The atoms in this aromatic ring were corrected manually a number of times during the refinement. By contrast, the atoms placed in the remainder of the density shown in Fig. 3 all refined quickly and without manual intervention from their positions on the original Sayre phase-refined map.

The final temperature factors for all the atoms (including solvent) range from 5·3 to 88·6 Å^2 with a mean value of 27·5 Å^2. For the protein the range is from 5·3 to 86·1 Å^2 (mean 21·1 Å^2) and for the solvent the range is from 11·1 to 88·6 Å^2 (mean 46·0 Å^2). Figs. 4 and 5 show for each molecule the mean temperature factors for atoms in the main chain and side chain of each residue. The distribution shown follows the expected pattern, being low for the main-chain residues forming helical segments (B9–B19 for

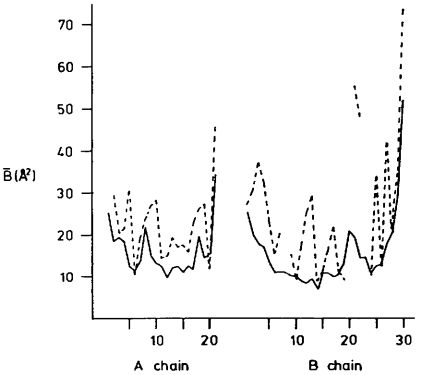

Fig. 4. The mean B values for molecule 1. For each residue the mean values are calculated for main-chain atoms (full line) and side-chain atoms (broken line) separately.

example) and larger at the ends of the chains. The large peaks for some side chains, *e.g.* A5 Gln molecule 2, possibly indicate errors in the coordinates.

We have obtained estimates of the accuracy of the refined structure from the inverse elements of the normal matrix, as well as from a plot of R factor against angle (Luzzati, 1952) (see Fig. 6) and from the estimated errors in the electron density (Cruickshank, 1949). These results are given in Table 5. For the complete structure, including solvent molecules, the mean value of the standard deviation estimated from the least-squares matrix is 0·09 Å. For the protein atoms only the mean e.s.d. is 0·07 Å and for the solvent atoms it is 0·14 Å. The average temperature factor is 27·5 Å^2 and, for this, Cruickshank's formula gives an overall e.s.d. of 0·06 Å. For the protein atoms only (mean B 21 Å^2) it gives an e.s.d. of 0·03 Å. The Luzzati plot (Fig. 6) gives a mean error of 0·13 Å for all the coordinates. There is a surprising consistency between these estimates. The standard deviations calculated from the least-squares matrix are the best estimate and a good reflection of the overall accuracy of the structure. The Luzzati plot gives an overestimate of the error since it assumes that the differences between $|F_o|$ and $|F_c|$ are caused only by error in the atom coordinates. We calculated Cruickshank's formula by assuming that all the atoms in the structure were carbon, that the unit-cell axes were orthogonal and that the errors were isotropic. In spite of these assumptions, the formula generally gives a reasonable estimate of the errors.

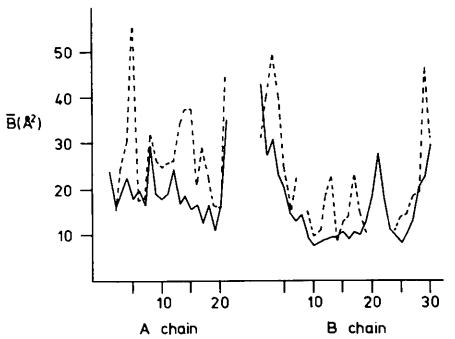

Fig. 5. The mean B values for molecule 2.

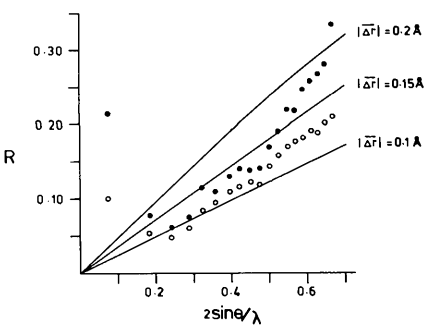

Fig. 6. The distribution of R value with angle, calculated for the final coordinates. The full circles are for all the data and the open circles for only those data used in the refinement (see text). Theoretical curves for different mean errors (Luzzati, 1952) are shown.

Radius of convergence

Fig. 7 shows, for the main-chain atoms only, the r.m.s. differences between various sets of coordinates and the final refined and regularized set of coordinates. The

Table 5. *Estimates of accuracy – coordinate standard deviations* (Å)

B range	Lsqs matrix	Cruickshank
6–10	0·033	0·007
11–15	0·041	0·014
16–20	0·052	0·024
21–25	0·063	0·038
26–30	0·080	0·057
31–35	0·099	0·080
36–40	0·114	0·108
41–45	0·125	0·140
46–50	0·158	0·178
51–55	0·158	0·222
56–60	0·204	0·272
All protein	0·07	0·03
All solvent	0·14	
Over all	0·09	0·06

(Luzzati plot gives $\bar{\Delta r}$ as 0·13 Å.)

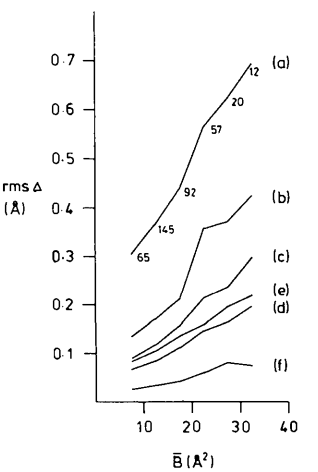

Fig. 7. The r.m.s. difference between the final coordinates of main-chain atoms and other sets of coordinates as a function of B value. The numbers refer to the number of atoms averaged at each point. (*a*) Coordinates read from the Sayre map; (*b*) coordinates after 14 cycles of refinement; (*c*) after 28 cycles; (*d*) after 37 cycles; (*e*) after 47 cycles and (*f*) after 58 cycles. All the sets of coordinates have been regularized except sets (*b*) and (*e*).

atoms have been grouped according to their refined temperature factors and the figure shows how an increase in temperature factor increases the initial positional error and decreases the rate of convergence. Of the 391 main-chain atoms positioned on the Sayre phase-refined map (Cutfield *et al.*, 1975) only 33 were manually adjusted during the course of the refinement. Fig. 7 then is a good illustration of the speed and ease with which the refinement procedure (least-squares shifts and regularization shifts) converged. The original coordinates differed from the refined coordinates by up to 0·7 Å, with the majority of the atoms differing by between 0·3 and 0·5 Å from their final positions. It is interesting to note that the regularized coordinates after 37 cycles of refinement (curve *d*) are closer to the final regularized coordinates than the unregularized coordinates after 47 cyles of refinement (curve *e*). This behaviour may be due to the tendency of the regularization procedure to produce a consistent structure for the main-chain atoms where the geometry is subject to a number of constraints (bond lengths, angles and planar peptide groups).

Comparison with difference-Fourier refinement

The least-squares refined coordinates were compared with the coordinates obtained after 14 cycles of difference-Fourier refinement (Dodson *et al.*, 1978). It was satisfying to find that for the bulk of the structure, particularly for the main-chain atoms in well defined regions, the two sets of coordinates agreed to within 0·2 Å. However, in both structures there were regions where incorrect interpretations (usually of weak electron density) had been made. As a result of the comparison, a 'best' set of coordinates was obtained and further refined by least squares. Details of this work will be presented elsewhere (Hodgkin *et al.*, to be published). Figs. 8 and 9 show for each residue of molecules 1 and 2, respectively, the r.m.s. differences between the least-squares coordinates and the refined 'best' set of coordinates. For all of the main-chain atoms, except four of the terminal residues, our refined coordinates agree, within the accuracy of the determination, with the coordinates of the refined 'best' structure. The worst agreements are with side-chain atoms. Without exception, these side chains have large temperature factors and the differences occurred mainly at those atoms with the largest temperature factors. The problem is essentially one of interpreting weak density – it is not a failure of the refinement method. This is nicely illustrated by comparing the least-squares refined coordinates with the 'best' coordinates for the disordered residues, B4 Gln of molecule 1 and B22 Arg of molecules 1 and 2. Fig. 10 shows, in projection, the positions of the two sets of atoms in each of the residues. The disorder in the Gln residue may be produced by a rotation of the side chain about the C_α—C_β bond, while in both Arg residues the disorder is produced by a rotation of the guanidino group about the C_γ—C_δ bond. In each case we had placed a water

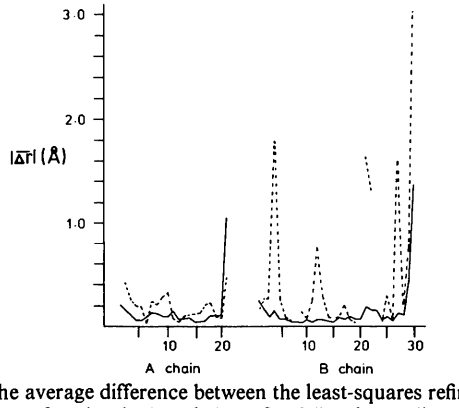

Fig. 8. The average difference between the least-squares refined coordinates of molecule 1 and the refined 'best' coordinates. The full line is for main-chain atoms, the broken line for side-chain.

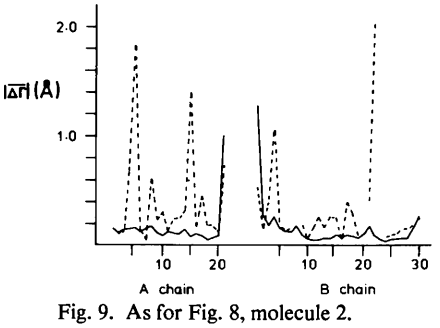

Fig. 9. As for Fig. 8, molecule 2.

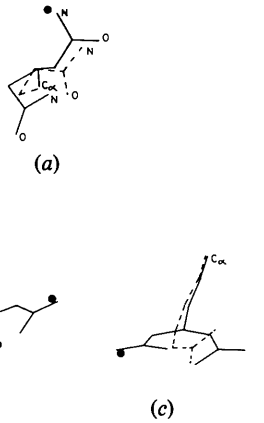

Fig. 10. The positions of the side-chain atoms of the disordered residues (*a*) B4 Gln of molecule 1, (*b*) B22 Arg of molecule 1, and (*c*) of molecule 2 projected on the *xy* plane. The full line shows the alternative conformation of the 'best' structure and the broken lines the conformation of the least-squares refined structure. Water molecules placed in the least-squares refinement are shown as closed circles.

molecule in the density of one of the disordered conformations. This accounted for some of the observed density and resulted in the refinement procedure moving the side-chain atoms to positions which are either close to correct positions or equidistant from two or more correct positions. For example, in B4 Gln, we placed a water molecule close (0·5 Å) to the position of one of the N_ϵ atoms of the disordered structure. The side chain was then incorrectly placed, but the atoms refined in such a way that N_ϵ was close (0·6 Å) to the correct C_δ, C_δ close (0·5 Å) to the correct C_γ, and C_γ close (0·8 Å) to the correct C_β (although the regularization procedure prevented too close an approach). The C_β atom moved to a position about midway between the C_γ and C_δ atoms of the alternative conformation, and O_ϵ moved to within 0·9 Å of the N_ϵ atom of the alternative conformation (see Fig. 10). In this way there is, overall, a quite good accounting for the observed weak density. A similar distribution of the least-squares refined atoms and solvent molecules about the correct positions of the disordered structures occurred in both B22 Arg residues (see Fig. 10). Often, large temperature factors indicated incorrectly placed atoms, *e.g.* C_β of B30 Ala in molecule 1 refined with a B of 77 Å^2. However, we did not find the temperature factor to be a reliable indication of wrongly placed atoms, since the protein molecule itself is quite flexible and large temperature factors ($B > 40$ Å^2) do exist. It is a considerable problem to distinguish a spuriously large temperature factor caused by an incorrectly placed atom from one due to truly large thermal motion. A possible solution to this problem is to calculate a series of difference-Fourier maps with a volume (say $\frac{1}{8}$) of the structure removed each time. This procedure has been used with considerable success in the difference Fourier refinement (Dodson *et al.*, 1978).

Conclusions

The fast Fourier least-squares algorithm has been shown to work effectively in the refinement of the structure of rhombohedral 2-Zn insulin. The considerable advantages of the program over conventional methods lie in its computational speed, wide radius of convergence and ease of use. Convergence is fast for well defined structure but slower for ill defined and highly disordered structure (atoms with large temperature factors). Our experience indicates that the method is capable of converging from large initial errors (*e.g.* 0·5 Å) without requiring manual intervention, but this convergence is slow. We feel that a much faster refinement could be achieved by making fuller use of the information in difference electron density maps and would recommend the following procedure as a guide for other protein-structure refinements.

(1) An $|F_o| - |F_c|$ map should be calculated with structure factors computed with the initial coordinates. Gross errors in coordinates should be corrected and any obvious shifts in coordinates or temperature factors should be applied manually.

(2) The initial refinement should be run on data to 2 Å resolution with the weights described earlier ($p = -1{\cdot}0$) applied to the ΔF terms. It would probably require about six cycles of refinement (four on coordinates, two on temperature factors) to produce convergence at this stage.

(3) The structure should be regularized with constraints on the positions according to the temperature factor of each atom. $|F_o| - |F_c|$ and $2|F_o| - |F_c|$ maps should be calculated and examined closely for gross errors, positions of solvent molecules, *etc.*

(4) Some further rounds of refinement (say three on coordinates, two on temperature factors) should be calculated with reduced weights ($p = -0{\cdot}5$) and all the data. After regularization, some further cycles of refinement should be run with either unit weights or $\sigma|F|$ weights, and all the data. At this stage the well defined structure should be almost refined. A series of $|F_o| - |F_c|$ maps with $\frac{1}{8}$ of the structure removed at a time should be calculated to check on the correctness of all the atomic coordinates.

(5) Hydrogen atoms should be included in the structure factor calculation and steps 3 and 4 repeated (with $p = 0$) until convergence is achieved.

We estimate that with this scheme and with close and detailed examination of difference Fourier maps, the refinement of a protein like insulin could be complete with about 30–40 cycles computed over a period of 6–12 months depending on the available manpower and computing facilities. Most of this time will be spent on interpreting difference Fourier maps to locate the poorly defined structure. If the diffraction data are measured at reduced temperatures, this will benefit considerably the speed and accuracy of the refinement. We emphasize again that the examination of difference Fourier maps is *essential* for a fast and complete convergence (as is the case in the refinement of small-molecule structures by least squares).

We thank Professor D. C. Hodgkin and Drs G. G. and E. J. Dodson for providing the insulin data and for their encouragement. The interest and advice of Drs R. L. Garwin, K. D. Hardman and D. Sayre provided constant stimulus for this work. NWI was the recipient of an IBM World Trade Research Fellowship.

References

ADAMS, M. J., BLUNDELL, T. L., DODSON, E. J., DODSON, G. G., VIJAYAN, M., BAKER, E. N., HARDING, M. M., HODGKIN, D. C., RIMMER, B. & SHEAT, S. (1969). *Nature (London)*, **224**, 491–495.

ADMAN, E. T., SIEKER, L. C. & JENSEN, L. H. (1975). *Acta Cryst* A**31**, S34.

AGARWAL, R. C. (1978). *Acta Cryst.* A**34**, 791–809.

BODE, W. & SCHWAGER, P. (1975). *J. Mol. Biol.* **98**, 693–717.
CHAMBERS, J. L. & STROUD, R. M. (1977). *Acta Cryst.* B**33**, 1824–1837.
COOLEY, J. W. & TUKEY, J. W. (1965). *Math. Comput.* **19**, 297–301.
CRUICKSHANK, D. W. J. (1949). *Acta Cryst.* **2**, 65–82.
CUTFIELD, J. F., DODSON, E. J., DODSON, G. G., HODGKIN, D. C., ISAACS, N. W., SAKABE, K. & SAKABE, N. (1975). *Acta Cryst.* A**31**, S21.
DEISENHOFER, J. & STEIGEMANN, W. (1975). *Acta Cryst.* B**31**, 238–250.
DIAMOND, R. (1971). *Acta Cryst.* A**27**, 436–452.
DIAMOND, R. (1974). *J. Mol. Biol.* **82**, 371–391.
DODSON, E. J., ISAACS, N. W. & ROLLETT, J. S. (1976). *Acta Cryst.* A**32**, 311–315.
DODSON, G. G., DODSON, E. J., HODGKIN, D. C. & VIJAYAN, M. (1978). To be published.
FREER, S. T., ALDEN, R. A., CARTER, C. W. & KRAUT, J. (1975). *J. Biol. Chem.* **250**, 46–54.
HODGKIN, D. C. (1974). *Proc. R. Soc. London Ser. A*, **338**, 251–275.
HUBER, R., KUKLA, D., BODE, W., SCHWAGER, P., BARTELS, K., DEISENHOFER, J. & STEIGEMANN, W. (1974). *J. Mol. Biol.* **89**, 73–101.
LUZZATI, V. (1952). *Acta Cryst.* **5**, 802–810.
MOEWS, P. C. & KRETSINGER, R. H. (1975). *J. Mol. Biol.* **91**, 201–228.
SAYRE, D. (1972). *Acta Cryst.* A**28**, 210–212.
SAYRE, D. (1974). *Acta Cryst.* A**30**, 180–184.
TAKANO, T. (1977*a*). *J. Mol. Biol.* **110**, 537–568.
TAKANO, T. (1977*b*). *J. Mol. Biol.* **110**, 569–584.
TEN EYCK, L. F. (1973). *Acta Cryst.* A**29**, 183–191.
WATENPAUGH, K. D., SIEKER, L. C., HERRIOT, J. R. & JENSEN, L. H. (1973). *Acta Cryst.* B**29**, 943–956.

Acta Cryst. (1978). A**34**, 791–809

A New Least-Squares Refinement Technique Based on the Fast Fourier Transform Algorithm

BY RAMESH C. AGARWAL*

IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, *USA*

(*Received* 25 *October* 1977; *accepted* 2 *May* 1978)

A new atomic-parameters least-squares refinement method is presented which makes use of the fast Fourier transform algorithm at all stages of the computation. For large structures, the amount of computation is almost proportional to the size of the structure making it very attractive for large biological structures such as proteins. In addition the method has a radius of convergence of approximately 0·75 Å making it applicable at a very early stage of the structure-determination process. The method has been tested on hypothetical as well as real structures. The method has been used to refine the structure of insulin at 1·5 Å resolution, barium beauvuricin complex at 1·2 Å resolution, and myoglobin at 2 Å resolution. Details of the method and brief summaries of its applications are presented in the paper.

Introduction

The conventional least-squares refinement method is prohibitively expensive for large structures such as proteins. In this paper, we present a new method which is inexpensive even for a very large structure.

The least-squares minimization

The process of least-squares minimization is discussed in *International Tables for X-ray Crystallography* (1959), Rollett (1965) and Lipson & Cochran (1966). It is being restated here to provide a uniform notation for the paper. In least-squares minimization, one minimizes the sum of squares of N error functions (differences between the observed and the calculated values) $E(s)$ with respect to (w.r.t.) R variables $u_r (N > R)$. The function to be minimized is called P and is given by

$$P = \tfrac{1}{2} \sum_{\mathbf{s}} E^2(\mathbf{s}). \qquad (1)$$

In (1), we have introduced the factor of $\frac{1}{2}$ to simplify the expressions for the derivatives. In normal least-squares minimization, the error function is approximated as a linear function of the variables as given by:

$$\Delta E(\mathbf{s}) = \sum_r \frac{\partial E(\mathbf{s})}{\partial u_r} \Delta u_r \qquad (2)$$

* Present address: Centre for Applied Research in Electronics, Indian Institute of Technology, New Delhi 110029, India.