Acta Cryst. (1980). A**36**, 559–572

Crystallographic Refinement of the Structure of Actinidin at 1·7 Å Resolution by Fast Fourier Least-Squares Methods

BY E. N. BAKER

Department of Chemistry, Biochemistry and Biophysics, Massey University, Palmerston North, New Zealand

AND E. J. DODSON

Chemistry Department, University of York, Heslington, York YO1 5DD, England

(*Received* 20 *August* 1979; *accepted* 24 *December* 1979)

Abstract

The structure of the proteolytic enzyme, actinidin, has been refined by fast Fourier least-squares methods [Agarwal (1978), *Acta Cryst.* A**34**, 791–809]. Atomic positions were refined independently by the least-squares program, with the whole protein structure being regularized at intervals. After an initial refinement phase with an overall temperature factor, B, only, individual isotropic B values for all atoms were also refined. Overall, the crystallographic R factor was reduced from 0·429 (for 14 800 reflections to 2·0 Å resolution) to 0·171 (for all 23 990 reflections between 10 and 1·7 Å resolution), with a final estimated accuracy in atomic positions of <0·1 Å. The final model comprises 1657 protein atoms, constrained close to standard geometry, and 163 solvent molecules, the latter identified using somewhat selective criteria. Most of the structure refined automatically with an average shift of 0·45 Å for main-chain atoms and 0·56 Å for side-chain atoms (maximum shift about 1·5 Å). Some larger shifts resulted from manual intervention. Groups of atoms with high B values, or which were not refining well, were removed at intervals for scrutiny in difference maps, and major corrections were made to the conformations of 16 side chains and two peptide units. One correction to the amino-acid sequence was made (Asp 86 → Glx 86) and disordered conformations were introduced for five side chains. The whole refinement was completed in three months.

1. Introduction

Most protein structures have been determined by fitting a molecular model to an electron-density map, calculated typically at a resolution of 2·5–3·0 Å. Such structures inevitably contain errors, partly attributable to errors in model building (*e.g.* in measuring the atomic coordinates), but more significantly to imperfections in parts of the original electron-density map – either lack of detail, ambiguities or distortions of the density (arising, for example, from errors in the experimentally determined phases). A number of protein structures have now been subjected to crystallographic refinement (for references, see below), and it is clear that this produces a substantial improvement in the structure. Not only is the accuracy of the protein structure enhanced, but so is the reliability with which other features (*e.g.* water molecules, ions, *etc.*) can be recognized.

We describe here the full refinement, at 1·7 Å resolution, of the structure of actinidin, a proteolytic enzyme of molecular weight 23 500, obtained from the fruit of the Chinese gooseberry, *Actinidia chinensis*. The amino-acid sequence of this enzyme has been determined (Carne & Moore, 1978) and a preliminary model for the three-dimensional structure obtained from an electron-density map at 2·8 Å resolution (Baker, 1977).

Refinements of protein structures to date have generally used either the difference-Fourier method (Freer, Alden, Carter & Kraut, 1975; Moews & Kretsinger, 1975; Adman, Sieker & Jensen, 1976; Chambers & Stroud, 1977) or the real-space refinement method of Diamond (1966, 1971, 1974) (Huber, Kukla, Bode, Schwager, Bartels, Deisenhofer & Steigemann, 1974; Deisenhofer & Steigemann, 1975; Bode & Schwager, 1975; Takano, 1977). While these methods have been applied most successfully, potentially the most powerful crystallographic refinement method (and the almost invariable choice for refining small structures) is that of reciprocal-space least-squares refinement. Conventional least-squares methods of this type have been used to refine parts of protein structures [*e.g.* the iron–sulphur clusters in ferredoxin (Adman, Sieker & Jensen, 1976)], but only once have they been used for a full protein refinement, *viz* in the refinement of the small (54-residue) electron-transfer protein rubredoxin (Watenpaugh, Sieker, Herriott & Jensen, 1973).

0567-7394/80/040559-14$01.00

Although the early refinement on rubredoxin was by means of difference-Fourier syntheses, the later least-squares refinement showed that the method could be used successfully for a protein structure, but that the computation time required might normally be excessive.

Recently, however, several new least-squares procedures have been developed, specifically for refinement of large molecular structures. Konnert (1976) has introduced a least-squares technique in which distance restraints are incorporated, so that not only is proper geometry maintained during refinement, but only a fraction of the available diffraction data need be used to achieve convergence. The speed of the refinement is thus greatly increased, and the method has been successfully applied in the refinement of the 108-residue carp calcium-binding protein. A similar least-squares method incorporating distance restraints on constrained groups has been developed by Sussmann and co-workers (Sussmann, Holbrook, Church & Kim, 1977; Sussmann, Holbrook, Warrant, Church & Kim, 1978) for refining protein and nucleic-acid structures.

A more radical variation of the least-squares technique has been developed by Agarwal (1978). In this approach, fast Fourier transforms are used at all possible stages of the computation so that computation time per cycle is drastically reduced. The method has been used (Isaacs & Agarwal, 1978) to refine the structure of rhombohedral 2-Zn insulin at 1·5 Å resolution. It is this method that we have chosen to refine the structure of actinidin, firstly because the experience with insulin suggested that it offered a remarkably fast and effective refinement method, when used judiciously; and secondly because we wished to assess its suitability for refining a larger structure (1650 protein atoms, compared with 813 for insulin) and one on which no prior refinement had been done. (The structure of insulin was partially refined by difference-Fourier methods before least-squares refinement was begun.)

The refined structure will be described separately; here details are included only where they illustrate aspects of the refinement. We have tried to include as many practical comments and illustrations on the refinement as possible, since we feel it provides such a remarkably fast, effective and easy-to-use method. In this case the whole refinement was completed in three months.

2. Experimental

(*a*) *Data collection*

The preparation and crystallization of actinidin have been described previously (Baker, 1974). Crystals are orthorhombic, cell dimensions $a = 78{\cdot}2$, $b = 81{\cdot}8$, $c = 33{\cdot}03$ Å, space group $P2_12_12_1$. Given the measured crystal density of 1·24 Mg m^{-3}, the weight per asymmetric unit is 39 000 daltons, corresponding to one protein molecule (~23 500 daltons) and approximately 850 water molecules.

All intensity data were collected on Hilger and Watts four-circle diffractometers. Four crystals were used to collect all the diffraction data to 1·7 Å resolution. In addition, a fifth crystal was used to recollect some sections of data for scaling purposes. Friedel pairs were measured for all shells except the outermost (covering the range 1·9 to 1·7 Å), for which long counting times were employed. An ω-scan technique was employed, with background counts measured on both sides of the reflection peak. Crystal alignment and deterioration were monitored by measuring the intensities of three standard reflections (chosen, where possible, from the resolution range being collected). These were measured after every 100 general reflections. The greatest fall-off in intensity for any of the four crystals was 25%.

For the innermost data (resolution range ∞ to 3·1 Å) the background correction for each reflection was taken from the individual background measurements for that reflection. For higher-resolution data (3·1 to 1·7 Å) a background averaging procedure, written by Dr P. E. Nixon, and similar to that suggested by Krieger, Chambers, Christoph, Stroud & Trus (1974), was used. Backgrounds were measured for relatively short times (5 s) on each side of the intensity scan. A subset of backgrounds was then taken, consisting of the background measurements for all reflections for which $I/B < K$ (where I = total counts for the scan through a reflection, B = total background estimated for that scan, K = ratio, typically 1·2 to 1·5). Thus only the weaker reflections (usually 30–50% of the reflections in a shell) were used so that the background measurements would not contain significant contributions from mis-set strong peaks. These backgrounds were then averaged as a function of θ, φ and χ, and the averaged values applied to all reflections in making background corrections. The reliability of the data was greatly improved by this procedure (as judged by comparison of symmetry-equivalent reflections and by comparison with photographs). This was particularly true of the weak reflections which made up a significant proportion of the data at high resolution.

For the whole data set, 72% of reflections had intensities greater than 2σ (where σ is the standard deviation derived from counting statistics); in the outermost shell (resolution 1·9 to 1·7 Å) this figure falls to 55%. No reflections were rejected, however. Those reflections for which the background-corrected intensity was negative (about 10% of reflections overall; 14% in the outermost shell) were given an intensity equal to $0{\cdot}5\sigma$ and were retained in the data set. Altogether the intensities of about 47 000 reflections were measured.

Intensities were corrected for radiation damage to the crystals. The fall-off in intensity of the standard reflections, and comparison of symmetry-equivalent reflections measured at the beginning and end of a shell were used to find the appropriate scaling corrections. The data from different crystals were scaled together using common reflections in overlapping shells. The intensities were corrected for Lorentz, polarization and empirical absorption factors (North, Phillips & Mathews, 1968). The overall merging R value for symmetry-equivalent reflections in the whole data set, R_m, was 0·056, where $R_m = \sum |I_l - \bar{I}| / \sum \bar{I}$ (I_l is the intensity of an individual measurement, $\bar{I}$ the mean value for that reflection, and the summation is over all measurements). After merging and scaling, the 47 000 measurements were reduced to a 1·7 Å data set of 24 157 independent reflections.

(*b*) *The least-squares refinement program*

The algorithms employed are basically those developed by Agarwal, and have been described in detail, with their theoretical basis (Agarwal, 1978). Structure factors are calculated by generating an atomic electron-density map, D_A, from the atomic coordinates, and then carrying out a fast Fourier transform on this map (Ten Eyck, 1973). Derivatives are obtained by calculating (again by fast Fourier transforms) gradient maps with coefficients of the form $-2\pi ihw(F_o - F_c) \exp i\alpha_c$, and integrating these with the atomic electron-density map, D_A. Separate maps are calculated to give the gradients in x, y, z or B. Shifts in the parameters are then obtained using the diagonal elements of the inverse least-squares matrix (off-diagonal terms are not calculated).

Other features of the program are that data are weighted as a function of $\sin\theta/\lambda$ [$W = (2 \sin\theta/\lambda)^p$] so that in the early stages of refinement lower-resolution data can be given higher weight (we set $p = -0{\cdot}5$ initially) and the weighting readily altered later (after the initial refinement phase we used $p = 0$, *i.e.* unit weights); data can be excluded on the basis of $\sigma(|F_o|)$ or the ratio of $|F_o|/|F_c|$; excessive shifts can be avoided by restricting the maximum shift to a pre-set multiple of the average (*e.g.* in the early stages of refinement no shifts greater than twice the average were allowed); and an optimum step size can be calculated (Agarwal, 1978) so that at the cost of only one more structure-factor calculation a better, scaled, set of shifts could be obtained. This latter criterion usually resulted in a comparable R factor being obtained for a set of shifts which were only 0·5–0·7 of those initially calculated. In the density-generating routine an additional increment (usually 15 Å^2) is made to atomic B values to allow a coarser grid to be used [this follows a suggestion by Ten Eyck (1977)]; the grid interval chosen was usually approximately one-third of the resolution of the data ($\simeq 0{\cdot}6$ Å for 1·7 Å data) for this and all FFT calculations.

The original program of Isaacs & Agarwal (1978) has been extensively modified so that it does not require large virtual memory facilities (in its present form it requires ~35K words of core storage for arrays *etc.*, and this could be reduced). Where possible all parts of the program have been generalized, so that symmetry-dependent operations are effected by reading in the standard equivalent positions for each space group. For example, when gradients are calculated for an atom close to the edge of the asymmetric unit, some of its density will overlap into the neighbouring asymmetric unit. Some of the contributions to the gradient must therefore be taken from the corresponding parts of a symmetry-related atom, and for this the gradient may have the opposite sign. Allowance is made for this, however, using the space-group symmetry. The density-generation routines are also general.

Fast Fourier transform routines used in the refinement program are space-group specific, however, and the appropriate routines must be inserted for each space group. (Versions of the refinement program are currently available for space groups $P1$, $P2_12_12_1$, $P4_12_12$, $P3$ and $P3_121$.) We have generally used the fast Fourier routines of Ten Eyck (1973), in some cases modified by G. Bricogne. In calculating the gradient maps, however, the standard fast Fourier routines in some cases must be modified, because pre-multiplication of the coefficients by $-ih$, $-ik$ or $-il$ (Agarwal, 1978) may change symmetry relationships for the space group. For example, for space group $P2_12_12_1$ coefficients $\bar{h}kz$ are generated using the relationship $T_{\bar{h}kz} = (-1)^k T_{hk\frac{1}{2}-z}$. When the coefficients have been premultiplied by $-il$ (*i.e.* for the z gradient) this relationship is changed, however, to $T_{\bar{h}kz} = (-1)(-1)^k T_{hk\frac{1}{2}-z}$.

Each xyz refinement cycle comprises at least two (usually three) structure-factor calculations and three gradient calculations. For a refinement cycle on actinidin, with 1820 atoms and 24 000 reflections and a grid, for FFT calculations, of 128 × 128 × 52 for the unit cell, timings were as follows:

on DEC-10 computer (University of York):

each structure-factor calculation	~ 3 min
each gradient calculation	~ 3 min
complete refinement cycle	~ 20 min.

Within each structure-factor calculation the times taken for the generation of the atomic electron-density map and for its Fourier transformation are about equal.

On a CDC Cyber 74-16 computer (University of Groningen) the times were even less, the time for a complete refinement cycle being approximately 10 min.

(*c*) *Regularization of the structure*

The least-squares refinement program applies no constraints to bond lengths, angles, *etc.* Therefore, it was necessary at intervals to regularize the structure, *i.e.* to constrain it to standard geometry. For this purpose, the model-fitting program (*MODELFIT*) of Dodson, Isaacs & Rollet (1976) was used. This simultaneously adjusts, by least squares, all bond lengths, angles and the planarities of groups. For actinidin the molecule was regularized in four sections (residues 1–55, 54–109, 108–163 and 162–218), because of core-storage restrictions.

The main advantages of this method of regularization are given below.

(i) It was fast and easy to use.

(ii) The strictness of regularization could be pre-set. In early stages of the refinement bond lengths were constrained to within 0·005 Å of standard values, and bond angles to within 1·5°; towards the end of the refinement these were relaxed somewhat so that bond lengths were within 0·015 Å and bond angles within 3·5°.

(iii) Atoms could be given individual standard deviations, so that more poorly defined parts of the structure (*e.g.* those parts with high *B* values) were constrained more rigorously. We based standard deviations on *B* values, using the expression suggested by Isaacs & Agarwal (1978), *viz* $\sigma = 0{\cdot}2\sqrt{B}/8\pi^2$. An alternative would be to use the standard deviations from the inverse least-squares matrix, but we did not, in fact, expand the program to calculate these until late in the refinement. One cautionary note is that the regularization takes no account of the hand of each residue. Thus if C^β of a Thr or Ile side chain was shifted a large amount in the least-squares refinement, relative to adjoining atoms, regularization could invert the configuration at C^β. This happened three times in the refinement of actinidin, twice with Thr side chains (the solution was then to interchange C^γ and O^γ), and once with an Ile (which was then rebuilt). The handedness at each asymmetric centre must be regularly checked, especially early in the refinement, when shifts are large.

(*d*) *The starting model*

The initial coordinate set for actinidin was that taken from the 2·8 Å resolution MIR* map (Baker, 1977). The mean figure of merit for the 5650 reflections used was 0·81. In general, the electron density was extremely well defined, but deficiencies noted at the time included the following: no density at all for the two final residues, Asn 219 and Asn 220, nor for C^β of Ala 93; very weak, discontinuous density for the side chains of Glu 21, Arg 63, Glu 87, Asp 97, Glu 114, Tyr 130 and Glu 191 (all external groups); ill-defined density for the side chains of Ser 42, Val 100, Gln 146 and Thr 171, allowing no convincing fit; unsatisfactory fits to apparently good density for Ile 110 and Ile 166; and unexplained extra density beyond C^β for Ala 101. Many of these residues had to be revised during refinement, and it is clearly advantageous to have a record of difficulties arising during the initial model building.

Coordinates were obtained by placing markers on the map in the Richards box, using the image of the model as a guide, but taking as first priority that atoms should be as well placed in the density as possible. Residue 93 was treated as Gly [it was noted that the corresponding residue in the related enzyme, papain (Drenth, Jansonius, Koekoek & Wolthers, 1971), is Gly] and residues 97 and 101 were transposed to become Ala and Asx. It was thought that these could have represented errors in the amino-acid-sequence determination. Poorly defined side chains noted above were placed in the best available density, giving a starting model consisting of 1650 protein atoms.

A structure-factor calculation based on this initial model gave a value of 0·371 for the residual, or *R* factor,* for the 5500 reflections of the 2·8 Å data set. Although a standard Wilson plot had suggested an overall *B* value of 15 Å^2 for the structure, it was felt that a value of 12 Å^2 was more realistic for the protein atoms, which should be better ordered than the solvent

* Multiple isomorphous replacement.

* Here, and elsewhere in this account, the *R* factor is defined as $R = \sum||F_o| - |F_c||/\sum|F_o|$, where $|F_o|$ and $|F_c|$ have been placed on the same scale. Apart from 170 inner reflections with spacings greater than 10 Å, which were omitted throughout the work, all reflections were included in the summation, and in refinement and structure-factor calculations. No other data were excluded on the basis of standard deviations or other criteria.

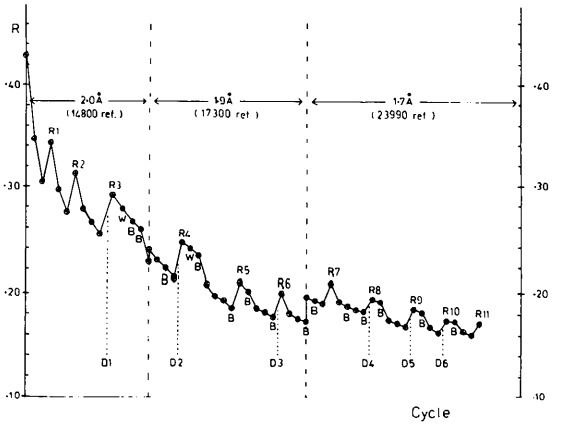


Fig. 1. A plot of *R* during refinement. Regularizations of the structure are marked *R*; *B* refinement cycles designated *B*; steps where the only change was the introduction of further solvent molecules *W*; unmarked points represent *xyz* refinement cycles. The points where difference maps were calculated are indicated (*D*1, *D*2, *etc.*). Stages where the data were extended to higher resolution are also shown.

atoms, and these were therefore all given an isotropic B value of 12 Å^2 in the structure-factor calculation.

Comparison of bond lengths in the initial model with standard values showed a r.m.s. deviation of 0·12 Å. When the geometry of the model was regularized, by *MODELFIT*, R dropped slightly, to 0·366. When 2·0 Å data became available and were included in the structure-factor calculation, R values were 0·435 and 0·429 for the unconstrained and regularized models respectively. The latter was therefore used as the starting point for least-squares refinement.

(e) *The general approach to refinement*

The progress of the refinement is summarized graphically in Fig. 1, where the crystallographic R factor is plotted against the least-squares refinement cycles. Altogether, 28 cycles of coordinate (*xyz*) refinement and 14 cycles of B refinement were carried out. No B refinement was done until the structure had been substantially improved by the initial *xyz* refinement. In the early stages of refinement the structure was regularized after every two cycles of *xyz* refinement; later, regularization was done after two or three cycles of *xyz* refinement and one or two cycles of (separate) B refinement. As far as possible, these automatic refinement procedures were allowed to continue without manual intervention; when manual intervention was required (to correct major errors *etc.*) difference maps were used with limited objectives in mind (not for major re-examination of the model).

Altogether, six difference electron-density maps were

Table 1. *Specifications of difference maps calculated during refinement*

Map number	Resolution	R*	Structure temporarily omitted in structure-factor calculation	Resulting changes to model	Atoms in model after map
*D*1	2·0 Å	0·254	None	(i) 95 and 151 peptides reoriented (ii) C^β put in for Ala 93 (iii) 16 side chains removed (21, 58, 63, 66, 75, 87, 94, 101, 114, 130, 142, 145, 146, 172, 191, 211) (iv) 76 solvent molecules included	1567 protein 76 solvent
*D*2	1·9 Å	0·212	(i) The 16 side chains left out after *D*1 (ii) 19 side chains (3, 4, 9, 15, 25, 33, 54, 59, 60, 78, 95, 100, 110, 120, 129, 138, 161, 171, 208) and 2 peptides (8, 142), with $B > 21$ Å^2	(i) New conformations for 9 side chains (33, 59, 78, 100, 110, 125, 138, 194, 208) (ii) 11 side chains left out after *D*1 now reincluded; 6 in new conformations (63, 66, 75, 94, 114, 211); 5 in original conformations (58, 130, 142, 145, 172) (iii) 21 solvent molecules added (iv) 2 oxygen atoms (half-weight) added to S^γ of Cys 25	1628 protein 97 solvent
*D*3	1·9 Å	0·175	(i) 5 side chains still left out since *D*1 (21, 87, 101, 146, 191) (ii) 27 side chains (18, 37, 39, 42, 44, 46, 54, 86, 94, 100, 105, 129, 135, 138, 139, 140, 142, 143, 148, 149, 153, 155, 160, 164, 165, 166, 195) and 19 peptides (21, 42, 43, 94, 95, 135, 139–143, 146, 148, 154, 155, 174, 187, 200, 201) with $B > 20$ Å^2, or large shifts on regularization (iii) Solvent molecules with $B > 30$ Å^2	(i) New conformations for 7 side chains (37, 39, 46, 54, 100, 166, 195) (ii) Side chains 21 and 101 added, and C^β for 87, 146 (iii) 5 solvent molecules omitted (iv) Disorder introduced for Ser 42	1637 protein 92 solvent
*D*4	1·7 Å	0·184	(i) 20 side chains with highest B values (3, 4, 21, 25, 58, 60, 63, 66, 87, 94, 97, 100, 114, 130, 138, 142, 145, 171, 172, 211) (ii) All solvent with $B > 20$ Å^2 (total 47)	(i) New conformations for 4 side chains (54, 100, 138, 171) (ii) Side chain 191 added (iii) 7 solvent molecules omitted, 42 new solvent molecules added	1642 protein 127 solvent
*D*5	1·7 Å	0·168	None	(i) 37 solvent molecules added (ii) Side chain 146 added (iii) Asp 86 changed to Glx 86 (iv) Disorder for Ser 9, Ser 44, Thr 129 (v) New conformation for side chain 138	1650 protein 164 solvent
*D*6	1·7 Å	0·163	(i) 7 residues with large shifts on regularization (58, 61, 105, 148, 164, 191, 211) (ii) Side chains 25, 162	(i) New conformation side chain 211 (ii) Disorder side chain 58 (iii) One O attached to Cys 25 repositioned	1657 protein 164 solvent

* The R factor is that at the end of the least-squares cycle preceding the map, and refers to an unconstrained model.

calculated during the refinement, and details of these are given in Table 1. In two cases (*D*1 and *D*5) the difference map was calculated following a structure-factor calculation in which all atoms of the current model were included. The purpose of these two maps was to look generally for significant, previously unrecognized errors in the structure and for solvent molecules. A further difference map of this type was calculated after refinement had been terminated, as a final check on the structure.

The other four difference maps (*D*2, *D*3, *D*4 and *D*6) were all calculated for the purposes of looking at the density of specific parts of the current model (peptide units, side chains, solvent molecules) which were suspected of being wrongly placed. Two criteria were used in selecting parts of the structure for special scrutiny: (i) atoms with high *B* values (or inconsistent *B* values within groups of atoms), and (ii) atoms which had experienced large shifts in the least-squares refinement cycles but had been moved back close to their original positions by the regularization procedures. Atoms in the first category may have high *B* values because they are wrongly placed, while atoms in the second category may be refining towards a new, possibly more correct position, but unable to remain there because of the constraints imposed by the adjacent structure. In either case, the suspect atoms (usually together with the other atoms in the side chain or whole residue) were omitted from the model. The density due to these selected parts of the structure then generally stood out clearly in the following difference map, making rebuilding easier.

Manual rebuilding of parts of the structure was done simply by plotting new positions for the atoms on to the electron-density sections (on paper), on a scale of 0·2 Å per mm. At a resolution of 1·7–1·9 Å interpretations were not often ambiguous. 'Labquip' molecular models were used as a guide to correct geometry, and although this meant that the geometry of rebuilt residues was seldom perfect, it was good enough for subsequent regularization or least-squares refinement. Rebuilding of specific residues was followed by either (i) regularization of the whole structure (with atoms of rebuilt structure given low standard deviations), before any further least-squares refinement, or (ii) insertion of the rebuilt, unregularized residues into the rest of the structure, followed by immediate least-squares refinement. The first procedure was used in the early part of refinement, when shifts tended to be large (giving large deviations from standard geometry); the second approach was satisfactory later in the refinement.

(*f*) *The course of the refinement*

The reduction in *R* during the refinement is shown in Fig. 1, along with other major steps (calculation of difference maps, introduction of higher-resolution data, *etc.*). Root-mean-square shifts in atomic positions during *xyz* refinement cycles or regularizations are shown in Fig. 2 (for *xyz* cycles, the shift shown is that actually applied after multiplying the calculated shift by the optimum step size, usually 0·6–0·8).

Initial refinement was very rapid. In the first two *xyz* cycles, *R* dropped from 0·429 to 0·305 for 2·0 Å data (14 800 reflections). (Our program was not fully tested at this stage, but we felt these results were encouraging!) Shifts were correspondingly large (average total shift 0·4 Å after two cycles) and produced considerable distortion from standard geometry (r.m.s. deviation of bond lengths from standard values 0·24 Å). Regularization of the structure (r.m.s. shift 0·22 Å) increased *R*, but only by about one-third of the decrease during the preceding two cycles; this pattern was repeated throughout the refinement. The net shift in atomic positions after two cycles of *xyz* refinement and one regularization was 0·32 Å. This initial phase of refinement (involving no *B* refinement and no manual intervention) was continued through a further five *xyz* cycles and two regularizations to an *R* (for a regularized structure) of 0·291. The total net shift in atomic positions was then 0·43 Å for main-chain atoms, 0·51 Å for side-chain atoms, on average.

It was clear that further refinement would require manual intervention to correct major errors, introduce solvent to the model, *etc.* The first difference map (*D*1), calculated at this stage, led to the omission of 16 side chains which occupied deep negative regions in the map (these were mostly side chains which had been poorly defined in the 2·8 Å isomorphous map) and to he inclusion of 76 solvent molecules (all regarded as water). No attempt was made to rebuild any of the protein structure but three obvious errors were corrected; the peptide-unit orientations of residues 95 and 151 (carbonyl groups rotated by about 60 and 120° respectively); and residue 93, which now had density for a β-carbon and became Ala 93. A further 57 peaks

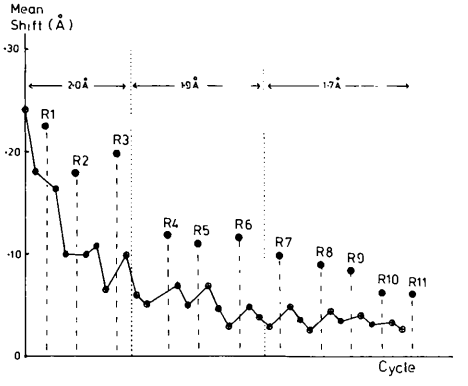


Fig. 2. Plot showing the mean shift in atomic positions in each *xyz* refinement cycle during the refinement. Mean shifts for regularizations (*R*) are also shown.

were noted as possible solvent molecules, but not included in the model, either because they were close to side chains whose orientations were in doubt, or because they were remote from other hydrogen-bonding groups. Only 21 of these were later confirmed as solvent peaks. In retrospect, other peaks in this map indicated errors in the structure which were corrected later, but their interpretation was not clear at this stage, and no time was spent on them.

When B refinement was begun, on this new model ($R = 0{\cdot}276$), some B values increased markedly, but the general effect was that most decreased to an average of about 10 Å^2. Some B values tended to become negative and since this was not physically reasonable, all were restrained to a minimum of 5·0 Å^2. In other cases, B values within groups of atoms showed considerable fluctuations. This may have resulted in part from the incomplete refinement of atomic positions, but most probably reflects the fact that 2·0 Å resolution is not high enough to give reliable individual B values. Later, as higher-resolution (1·7 Å) data were included, B values were greatly improved and none showed any tendency to drop below 4·0 Å^2.

Extension of the data to 1·9 Å (17 300 reflections) caused a temporary increase in R of 0·015 (due to initial poor agreement of the new high-resolution data) but had little effect on the size of shifts. Further refinement mostly depended on identification and correction of errors in the protein structure and addition or deletion of solvent. Most corrections were made following difference map $D2$, prior to which all groups containing atoms with $B > 21$ Å^2 were omitted from the model. New conformations were found for 10 out of 20 of the side chains so omitted. Clearly interpretable density was also seen for 11 of the 16 side chains omitted earlier (of the remaining five, two were reincluded after difference map $D3$, one after $D4$, one after $D5$, but no density was ever seen beyond C^β for the fifth, Glu 87). Residues with highest B values were re-checked later in the refinement (difference maps $D3$ and $D4$) but were almost all confirmed in their previous conformations.

Comparison of coordinate sets, to pin-point residues which received large shifts on least-squares refinement but were returned to their old positions on regularization, led us to check other residues in the later stages of refinement ($R < 0{\cdot}24$). Most of these had B values which were relatively low (<20 Å^2), but often inconsistent, (*e.g.* $C^{\gamma 1}$ greater than $C^{\delta 1}$ for Ile side chains, or higher B values in the middle of a Lys side chain). Out of 13 such side chains with average shifts >0·25 Å on regularization, six were found (in difference map $D3$) to have wrong conformations, and one to be disordered. Two more were found to be disordered later in the refinement, and one (residue 86) was identified as a possible sequence error and changed from Asp to Glx.

Five side chains were rebuilt more than once (residues 54, 100, 138, 171 and 211), and it was not until 1·7 Å data were included in the refinement that the correct conformations for these side chains became clear. Checks were also made on several other residues which were of particular interest (*e.g.* the *cis*-Pro residue at 153, and the catalytically-important Cys 25 and His 162) by temporarily omitting them from the model before a difference-map calculation.

Two 'bumps' on the peak for the sulphur atom of Cys 25 were taken as oxygen atoms, arising from the partial oxidation of the sulphydryl group (probably to SO_2^-). One was very clear, occupying the 'oxyanion hole' for substrate binding, at hydrogen-bond distance from the main-chain NH of Cys 25, and the side-chain amide of Gln 19. The other was less certain, and there may be some slight amount of disorder. Both were given an occupancy of 0·5, consistent with a degree of oxidation of about 50%.

The solvent structure was reviewed several times. Before difference map $D3$, solvent molecules with $B > 30$ Å^2 were omitted from the model (with 13 out of 18 then being returned as definite). Before difference map $D4$, all solvent molecules with $B > 20$ Å^2 were temporarily omitted, with 40 out of 47 then being returned. Further solvent molecules were included as they appeared in difference maps, but only if they had persisted through several such maps, and were at reasonable hydrogen-bonding distance from other atoms.

During most of the refinement, shifts were restricted to a maximum of three times the average shift for a least-squares cycle, but this ratio was increased to five towards the end. A larger ratio is needed for limiting the shifts in B values as the refinement converges, since some atoms genuinely have very high B values (50–70 Å^2) and may otherwise take a long time to reach there.

(g) *Convergence of the refinement*

Refinement was terminated when R, for 23 990 reflections between 10 and 1·7 Å resolution, was 0·171 for a regularized structure (r.m.s. deviations from standard values 0·014 Å for bond lengths, 3·2° for bond angles). Shifts on least-squares refinement had become very small (average 0·03–0·04 Å) for atomic positions and B values were almost constant. Least-squares refinement on this regularized model reduced R to 0·158 after three cycles, but regularization then returned R to 0·171, and there was no significant difference between the two regularized structures (average difference 0·016 Å). Thus convergence appeared to have been reached.

Comparison of the constrained ($R = 0{\cdot}171$) and unconstrained ($R = 0{\cdot}158$) models (Fig. 3) showed an average difference between atomic positions of 0·05 Å, with 95% of main-chain atoms and 87% of side-chain

atoms less than 0·1 Å apart. Clearly most of the protein structure showed little tendency to refine away from the regularized model. In the unconstrained model the r.m.s. deviation from standard bond lengths was 0·07 Å (maximum deviation ~0·4 Å) and from standard bond angles 5·0° (maximum ~15°).

For eight residues there were still atoms which shifted more than 0·2 Å on refinement (but returned on regularization). Of these, five had high B values, with two (Glu 21 and Met 211) also showing obvious, but not readily interpretable, signs of disorder in difference maps. The remaining three (Gln 105, Ala 148, Ile 164) have moderate B values. Although improved orientations could not be found, these residues remain as possible errors in the model.

A final difference map was relatively featureless. No peak was greater than 0·5 e/Å^3 in height, and only ten greater than 0·4 e/Å^3. Of those between 0·3 and 0·5 e/Å^3, about 80–100 could be interpreted as low-occupancy solvent molecules, from their positions with respect to other solvent and protein atoms, but at this low level of density they have not been included in the model. Peaks of 0·3 to 0·5 e/Å^3 associated with the side chains of Cys 25, Leu 53, Asn 61, Gln 94, Ala 101, Ala 148, Ile 164, Val 165, Tyr 169 and Met 211, and the main chain between 173–174, may represent slight disorder in these groups, or at least inadequate description of their thermal motion. There are a few negative peaks of less than −0·4 e/Å^3 (mostly where side chains have high B values, or where there is some disorder) but none less than −0·5 e/Å^3.

The final (regularized) model comprises 1659 protein atoms (24 of partial occupancy) and 163 water molecules. The only parts of the protein not included are the side chain of Glu 87 (no density beyond C$^\beta$) and residues Asn 219 and Asn 220 (never seen).

3. Discussion

(*a*) *Reliability and accuracy*

The most obvious crystallographic indication of the success of the refinement is the reduction of the R factor from a starting value of 0·43 (for 2·0 Å data) to a final value of 0·171 (for 1·7 Å data). Thus the final model for the structure gives a high level of agreement with the observed X-ray data. (Although small-molecule structure analyses commonly lead to R values less than 0·10, few protein structures have been refined to R values less than 0·20.)

The value of R depends in part on the resolution of the data, on the amount of data omitted from the calculation, and on the degree of stereochemical constraint imposed on the structure (see Table 2). Apart from 170 reflections with spacings >10 Å (which are strongly influenced by the solvent continuum), we have included all data in our calculations, including those reflections usually classed as 'unobserved'. The agreement is poor for these weak data (see Fig. 4*b*), and if reflections for which the intensity $I_{hkl} < 2\sigma_I$ are omitted, as is frequently done, the final R value is reduced to 0·152. Nevertheless, we feel it is better to include these weak data, approximately 4200 reflections, since the higher ratio of observations to parameters should then give a more reliable structure, with lower standard deviations.

Table 2. *Variation of R values*

(i) R values for different models and data sets

Model	Data set	N*	R
Regularized	1·7 Å, all data	23 990	0·171
Regularized	1·7 Å, omitting $I < 2\sigma$	19 724	0·152
Unconstrained	1·7 Å, all data	23 990	0·158

(ii) Variation of R with resolution (for regularized model)

4 sin$^2\theta/\lambda^2$	N (all 1·7 Å data)	R (all 1·7 Å data)	N (1·7 Å data, omitting $I < 2\sigma$)	R (1·7 Å data, omitting $I < 2\sigma$)
0–0·05	1313	0·191	1295	0·188
0·05–0·10	2483	0·136	2436	0·130
0·10–0·15	3160	0·160	2843	0·147
0·15–0·20	3703	0·165	3097	0·144
0·20–0·25	4143	0·173	3465	0·150
0·25–0·30	4585	0·197	3475	0·159
0·30–0·35	4603	0·236	3113	0·180

(iii) Estimated standard deviations in atomic positions (Å)

	B = 0–5	5–10	10–15	15–20	>20 Å^2	Overall
From least-squares matrix	0·041	0·050	0·063	0·073	0·081	0·055
From shifts in final refinement cycles	0·040	0·048	0·082	0·126	0·151	0·071
From variation of R with resolution	–	–	–	–	–	0·10 (maximum)

* N = Number of reflections.

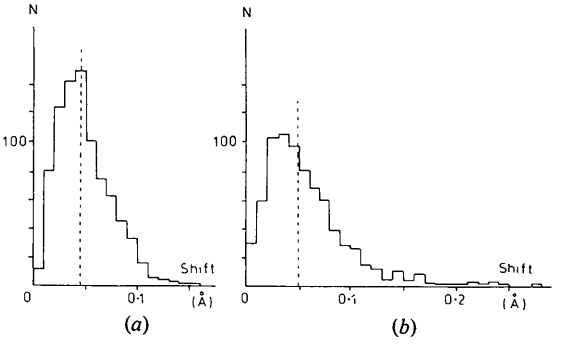


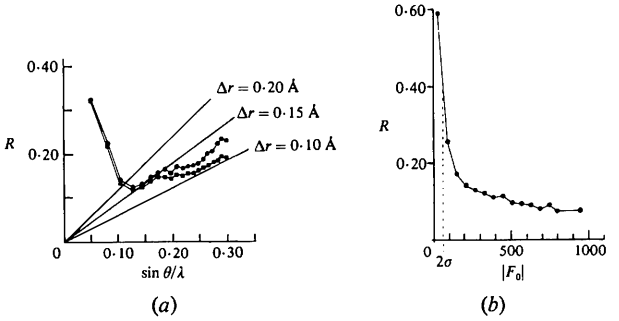
Fig. 3. Histogram showing the differences between the final regularized model for actinidin and an unconstrained model, at the end of refinement. In (*a*) the distribution for main-chain atoms is shown (median difference 0·045 Å). In (*b*) that for side-chain atoms is shown (median difference 0·049 Å).

The final model is constrained to a degree that ensures that no part of the structure deviates unacceptably from standard geometry [r.m.s. deviation 0·014 Å in bond lengths (maximum ~0·04 Å) and 3·2° in bond angles (maximum ~10°)]. If the constraints are relaxed, R is reduced, but at the expense of some unlikely distortions in the structure.

The accuracy of the model has been estimated in several ways (see Table 2). The standard deviations in atomic positions can be obtained from the inverse elements of the normal matrix; these vary from 0·04 Å (for atoms with low B values) to 0·08 Å (for atoms with high B values), with an average value of 0·055 Å. A similar value can be deduced from consideration of shifts in the final refinement cycles. If refinement of the final regularized model is allowed to continue until convergence, the average shift is 0·065 Å, and the resulting unconstrained model shows a r.m.s. deviation in bond lengths of 0·07 Å.

An upper level for the positional error can be obtained by plotting R as a function of resolution (Luzzatti, 1952). The variation of R with resolution is shown in Fig. 4(*a*), together with the theoretical plots for non-centric data, given by Luzzatti. Since the latter assume that the only errors are those in the model it is probably more realistic to use the data set from which weak reflections ($I < 2\sigma$) have been omitted. The plot for this data suggests a maximum error in atomic positions of 0·10 Å. (Note that this may still be a slight overestimate since about 15% of the data is centric.)

One independent check on the accuracy is given by the S—S bonds in the three disulphide bridges. These were never regularized during refinement (each Cys residue being treated as independent), yet the S—S bonds at the end of the refinement are closely similar (1·97, 2·01 and 1·96 Å). These values are slightly, but not significantly, shorter than the S—S bond lengths of 2·00–2·10 Å found in small organic disulphides (Shefter, 1970). The C—S—S angles (also not regularized) range from 101·0 to 108·3° (*cf.* 98 to 106° in small molecules) and the C—S—S—C dihedral angles are 90·8, 84·5 and 96·0° (expected value about 90°).



(*a*) (*b*)

Fig. 4. Plots showing the variation of the crystallographic R factor with (*a*) resolution and (*b*) the observed structure amplitudes, $|F_o|$. In (*a*) points ⊕ (upper curve) refer to the whole 1·7 Å data set, while points ■ (lower curve) refer to 1·7 Å data with weak data ($I < 2\sigma$) removed. Lines are also drawn to show the theoretical variation of R with resolution for different maximum errors in atomic positions (for non-centric data). In (*b*) the approximate cut-off for reflections with $I < 2\sigma$ is indicated.

(*b*) *Improvement of the protein structure*

As has been observed in other protein refinements, the definition of the electron density is greatly improved. Fig. 5 shows the improvement in several contrasting parts of the structure, *viz* a section of the polypeptide chain (167 to 171) which had been generally well resolved in the 2·8 Å map, and two side chains (Arg 63 and Glu 191) whose density was weak and discontinuous originally. The side chain of Glu 191 is, in fact, one of the three weakest-density groups in the final model.

Comparison of the atomic positions (for protein atoms) before and after refinement shows that quite large shifts have occurred, despite the fact that the initial structure was apparently better than most unrefined protein structures (as judged by the initial R value). Histograms showing the average shifts for all residues are given in Fig. 6. The median shift for all main-chain atoms is 0·45 Å, while that for side-chain atoms is 0·56 Å. Only two peptide units required a major reorientation during refinement, although many did change significantly. On the other hand, completely new conformations were found for 16 side chains (Thr 33, Ile 37, Lys 39, Thr 59, Gln 75, Ile 78, Gln 94, Val 100, Ile 110, Gln 125, Ile 166, Thr 171,

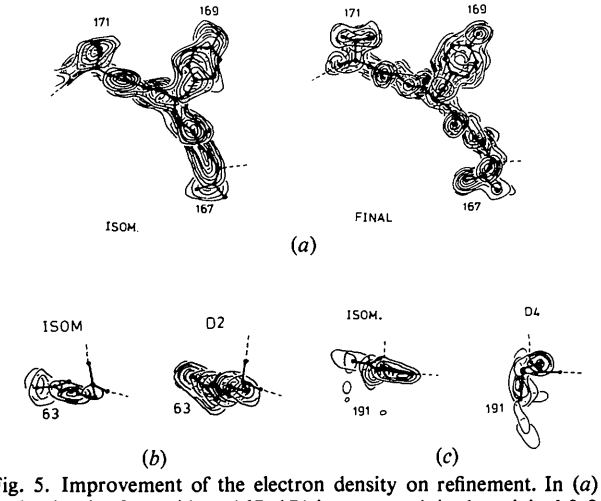
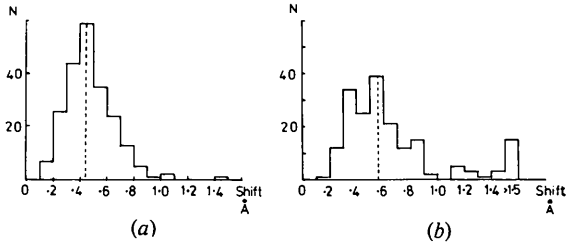


Fig. 5. Improvement of the electron density on refinement. In (*a*) the density for residues 167–171 is compared, in the original 2·8 Å isomorphous map, and in a final difference map (after omission of 167–171 from the model). In addition to the sharpening of the electron density, a corrected orientation for Thr 171, for which no satisfactory fit was originally achieved, can be seen. In (*b*) and (*c*) the densities for the side chains of Arg 63 and Glu 191 are shown, comparing the isomorphous density with that in later difference maps. Both had very poor density originally, were omitted during early stages of refinement and reincluded in the model as their density improved.

Glu 191, Ile 208 and Met 211), while in a further five residues the orientations of carboxyl or guanidinium groups were substantially altered. Ile side-chain configurations appear to have been particularly susceptible to misinterpretation.

The general improvement of the main-chain conformation is demonstrated graphically in Fig. 7, where the main-chain torsional angles are plotted in the familiar Ramachandran conformational map (Ramakrishnan & Ramachandran, 1965). In going from the initial, unrefined structure to the final, refined structure, there is a pronounced clustering of the residues in, or just outside the 'allowed' regions; in the refined structure, only two residues, other than Gly, lie significantly outside these 'allowed' regions.

Non-bonded contacts in the molecule are few (although no restrictions were placed on them during refinement). There are only 28 such contacts <3·2 Å, and of these only six are <3·1 Å. Amongst the latter, three (2·88, 3·06, 3·09 Å) are contacts between C and O atoms, and could represent C—H···O interactions, two (3·06, 3·08 Å) involve the side chain of Ile 164, about whose structure there remains some doubt, while the remaining one (2·98 Å) involves a disordered side chain (Arg 58).



(a) (b)

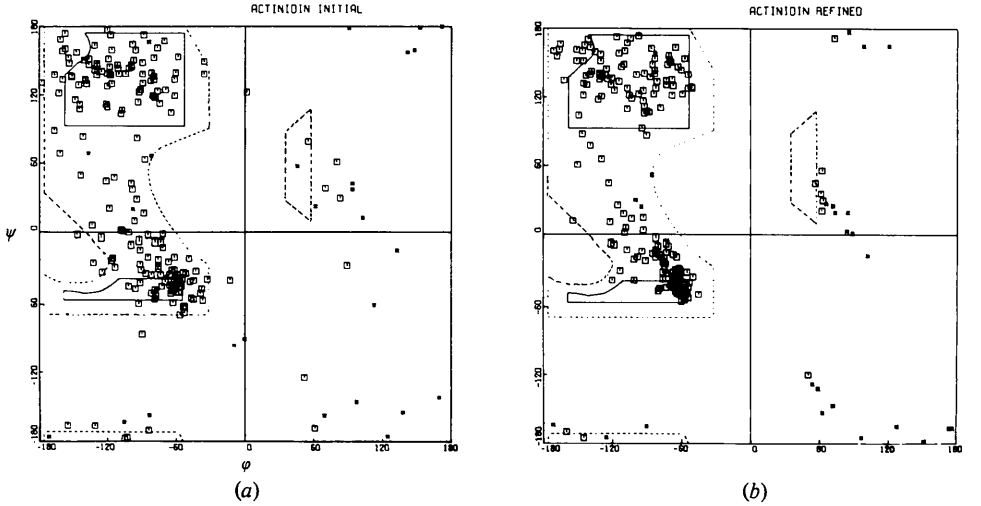
Fig. 6. Histograms showing the distribution of shifts resulting from refinement of the protein structure. In (*a*) the average shifts for the main-chain atoms of all residues are plotted, while in (*b*) the average shifts for all side chains are plotted. For 16 side chains the average shift is >1·5 Å.

(c) *Amino-acid-sequence changes*

The improvement of the phases (and hence the electron density) resulting from refinement has enabled several workers to obtain substantial amounts of amino-acid-sequence information, where little or no chemical evidence was available (Watenpaugh, Sieker, Herriott & Jensen, 1973; Anderson, McDonald & Steitz, 1978; Anderson, Stenkamp & Steitz, 1978). Since for actinidin the entire amino-acid sequence was already known (Carne & Moore, 1978), we did not set out to re-examine it. Nevertheless, as a result of the refinement, apparent discrepancies between the chemically-determined sequence and the original 2·8 Å map have been cleared up, and we have also been led to make one (and possibly two) changes to the sequence.

There were originally three apparent disagreements between the amino-acid sequence and the 2·8 Å isomorphous map. In each case the chemical assignment was confirmed on refinement; residue 93 (which appeared as Gly originally) quickly developed a β-carbon atom to become Ala 93 (see Fig. 8); residue 97 gradually developed weak density beyond C^{β} to become Asp 97 instead of Ala; while residue 101 lost its density beyond C^{β} to become Ala, as in the sequence.

We have, however, changed the identity of residue 86, from Asp, as in the chemically determined sequence, to Glx. Although the side-chain density was strong, and B values reasonably low (5 to 16 Å^2), side-chain atoms (especially C^{β} and the carboxyl oxygen atoms) consistently received large shifts on least-squares refinement, but were then pulled back to their former positions on regularization. Attempts to fit



(a) (b)

Fig. 7. Ramachandran plots of the conformational angles, φ and ψ, for (*a*) the initial, unrefined structure and (*b*) the final, refined model. Gly residues are shown by *, others by □.

an Asp residue in alternative orientations were unsuccessful. Finally, at a late stage of refinement (R = 0·186 for 1·7 Å data) we were forced to conclude that residue 86 was Glx rather than Asp. A difference map showed positive peaks on either side of C^{β}, and again beyond $O^{\delta 1}$ and $O^{\delta 2}$, consistent with a slight change in the position of C^{β} and a lengthening of the side chain by one carbon atom. As shown in Fig. 8, the new structure (as Glx) fitted perfectly, and the atoms were hardly moved by subsequent refinement or regularization.

One further possible sequence error concerns Ser 42. At present it has been treated as having a disordered Ser side chain (see Fig. 12 and later discussion), but it would fit equally well as Thr with full-weight atoms on both peaks. In fact, residual positive density for O^{γ} and $O^{\gamma 1}$ in the final difference map, in spite of low B values (4·6 and 6·5 Å^2) may be more consistent with Thr. The preceding residue, 41, is Thr, making a sequencing error possible, and it may also be significant that the corresponding residue in papain is Thr 42.

(*d*) *Solvent structure*

The original 2·8 Å electron-density map contained a number of peaks which were suggestive of water molecules or ions. Some appeared particularly convincing (*e.g.* in the active-site region, in pockets on the outside surface of the protein, and in an internal region between the two domains of the molecule, adjacent to a group of four charged side chains, Lys 17, Lys 181, Glu 35 and Glu 50). Coordinates were recorded for 29 of them, but they were not included in the initial model, and in fact no solvent was added until the initial phase of refinement had been completed (and R reduced to 0·29 for 2·0 Å data). In retrospect, 18 of the 29 solvent molecules noted originally were confirmed at this stage, but one very prominent peak near the active site (2·5 Å from the carboxyl group of Asp 161, and about 3 Å away from any heavy-atom site) disappeared completely. The positions of other peaks were substantially displaced, and it is clear that deductions about solvent structure from an isomorphously phased electron-density map of this sort of resolution (2·5–3·0 Å) are likely to be rather unreliable (see for example Fig. 9).

Some solvent molecules included in the model during refinement may have been ammonium ions; none appeared to have sufficiently striking size or density to be labelled as phosphate or sulphate ions. In the present model, therefore, all solvent molecules are regarded as water. All were located from difference maps, but were only included in the model if they were within approximate hydrogen-bonding distance (2·5–3·2 Å) of appropriate atoms in the protein, or other, previously identified, water molecules. The positions (*xyz* coordinates) and B values of the water molecules were refined as for the protein structure, but with no constraints (*i.e.* there was no attempt to limit their approach to other atoms, or to each other, on refinement). Despite this, and the quite large shifts seen for some (maximum 0·7 Å, average shift about 0·25 Å), only three out of 164 in the final model make approaches <2·5 Å to other atoms, and these short contacts (2·18, 2·31 and 2·34 Å) all involve water molecules with high B values. The average nearest-neighbour distance is 2·76 Å with a standard deviation (from the spread of values) of 0·2 Å.

The distribution of B values for solvent molecules is shown in Fig. 10(*b*). 48 out of 164 have $B < 20$ Å^2, *i.e.* are well ordered and apparently firmly bound to the protein. Apart from six which occupy intermolecular crevices, all of these tightly bound solvent molecules are either internal (total 17) or in pockets formed by polar groups on the surface of the protein, and should be regarded as an integral part of the structure. Eight of the 17 internal solvent molecules form a buried network extending below the active site and along the interface between the two halves of the molecule.

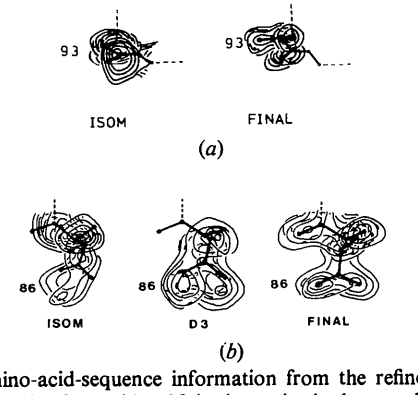


Fig. 8. Amino-acid-sequence information from the refinement. In (*a*) the density for residue 93 is shown in the isomorphous map (where it was built as Gly) and in a final difference map, when it was clearly Ala, as in the chemically determined sequence. In (*b*) the density for residue 86 is shown in the isomorphous map and in an intermediate difference map, *D*3 (where it was fitted as Asp, as in the chemically determined sequence); and in a final difference map when it had been changed to Glx.

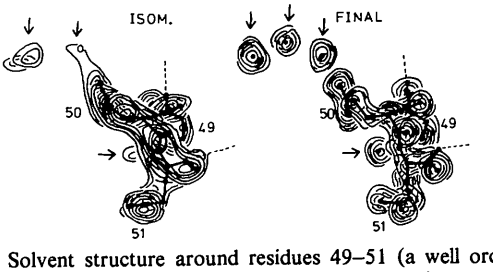


Fig. 9. Solvent structure around residues 49–51 (a well ordered, internal part of the protein structure), showing the improvement on refinement. In the original isomorphous map (left) there appear to be three water molecules (arrowed), two of them adjacent to the carboxyl group of Glu 50. After refinement, a final difference map (calculated after the omission of residues 49–51 and all solvent molecules from the model), clearly shows four water molecules, three of them adjacent to Glu 50.

The remaining solvent molecules mostly interact with polar groups on the protein surface. The absence of B values >50 Å^2 for solvent reflects the rather selective criteria used in ascribing peaks in difference maps to solvent. Although peaks can be seen in the final difference map which are likely to represent further genuine, if poorly ordered, water molecules (the peaks being of low density, $<0{\cdot}5$ e/Å^3), we have chosen only to include the more certain solvent molecules. Much of the (less-ordered) solvent in the crystal therefore remains unaccounted for.

(*e*) *Significance of individual B values*

For a protein structure, the B values of the atoms cover both thermal vibrations and small-scale disorder in atomic positions. If the inclusion of individual B values for atoms, as parameters in the refinement, is justified, the B values should reflect the degree to which atoms are free to move. We should expect main-chain atoms, and atoms of buried side chains to show consistently smaller B values than those on the protein surface, and of course the solvent structure.

Histograms showing the distribution of B values for (*a*) protein atoms and (*b*) water molecules are presented in Fig. 10. 70% of protein atoms have B values <12 Å^2, with an average value of 9·4 Å^2. A small group of atoms have high B values (30–60 Å^2), all of these being atoms of side chains projecting out from the molecule into the external solution, and presumably with a high degree of freedom. In some parts of the structure the B values are remarkably consistent. In the central helix, residues 25–42, only five main-chain atoms (out of 72) and seven side-chain atoms (out of 61) have B values >10 Å^2. A similar effect is seen in Fig. 11 where B values of main-chain atoms for all residues along the polypeptide chain are plotted. Those parts of the chain constrained by their involvement in helical structure or β structure clearly have lower B values than those parts which form more loosely held external loops.

An indication of the consistency of B values within groups can be gained from Table 3, where those of all Lys and Phe side chains (as examples) are listed. B values within the aromatic rings of the Phe residues are generally consistent. Among the Lys side chains, Lys 17 and Lys 181 are internal, involved in interactions with the carboxyl groups of Glu 35 and Glu 50, and a cluster of internal water molecules; their B values are, accordingly, low. Lys 39 is confined in a surface pocket and its terminal N atom forms three hydrogen bonds; again the B values are low. Lys 106 and 217 are more exposed (although in both cases N^{ζ} forms several hydrogen bonds), and have consistently higher B values. Lys 145 projects right out into the external solution and B values increase to 50 Å^2 at the end of the side chain. It is generally true that B values increase towards the end of the extended side chains; in Ile side chains $C^{\delta 1}$ usually has a higher B value than $C^{\gamma 1}$; and in peptide units the carbonyl oxygen atom almost invariably has a higher B value than its associated carbon atom. Thus we feel we can attach considerable physical significance to the values.

(*f*) *Disorder*

We were careful not to assume disordered conformations too early in the refinement; in fact, several residues which at first appeared to be disordered later turned out to have only a single conformation. In a few other cases, however, behaviour on refinement was such that disordered conformations had to be introduced.

An example is Ser 42, which had distorted, difficult-to-fit density in the isomorphous map (see Fig. 12*c*). On refinement C^{β} always received large shifts (0·4–0·5 Å) but was then returned to its original position on regularization. On leaving the side chain out of the model, a difference map showed density adjacent to C^{β} consistent with an alternative, lower-occupancy position for O^{γ} (achieved by rotation about the C^{α}–C^{β} bond). C^{β} apparently tended to refine towards this

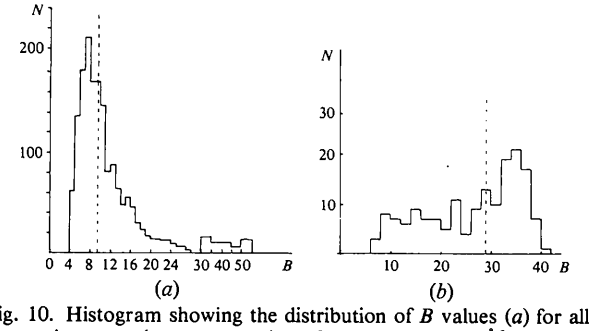


Fig. 10. Histogram showing the distribution of B values (*a*) for all protein atoms (note contraction of scale for $B > 30$ Å^2), and (*b*) for all solvent atoms.

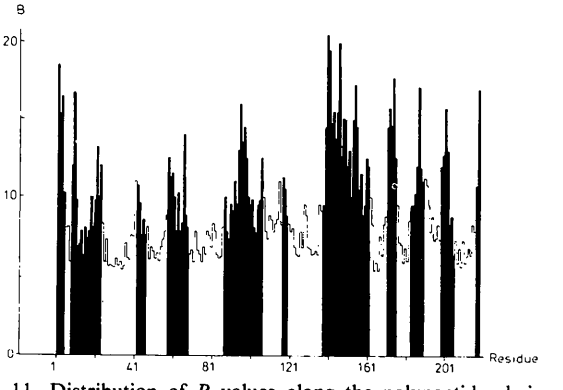


Fig. 11. Distribution of B values along the polypeptide chain of actinidin. For each residue the B value plotted is the average of those for C^{α}, C, O, N atoms. Helical sections of the chain and stretches involved in β structure (shown unshaded) have noticeably lower B values than external loops and less-ordered parts of the structure (shown black).

Table 3. *Final B values for all Lys and Phe side chains*

Lys	C^{β}	C^{γ}	C^{δ}	C^{ϵ}	N^{ζ}		
17	6·6	7·8	6·0	5·8	7·6		
39	8·0	5·2	8·7	7·1	9·2		
106	10·2	12·4	13·4	24·2	21·7		
145	16·3	22·8	33·8	50·0	48·3		
181	4·5	7·2	9·8	12·8	8·0		
217	12·7	16·4	18·7	15·4	14·0		
Phe	C^{β}	C^{γ}	$C^{\delta 1}$	$C^{\epsilon 1}$	C^{ζ}	$C^{\epsilon 2}$	$C^{\delta 2}$
28	6·8	5·4	8·4	9·5	8·5	6·1	6·3
74	5·0	5·8	5·2	8·7	6·3	5·3	8·5
76	4·8	6·7	9·1	9·0	7·0	9·2	9·6
144	9·9	9·0	13·2	15·6	13·2	18·9	13·7
152	12·5	10·4	10·0	9·3	15·2	12·8	11·3

position but the constraints of the adjacent structure pulled it back. [Note that an alternative interpretation would be to change residue 42 to Thr (see earlier discussion) but this cannot be resolved from the crystallographic evidence alone.] Similar persistent peaks adjacent to side-chain atoms have led to the introduction of disorder in Ser 9, Ser 44, Thr 129 and Arg 58.

(*g*) *Appraisal of the refinement methods*

Our experience with the fast Fourier least-squares refinement method is that it is remarkably fast and effective. The whole refinement, to a level of accuracy matched so far by very few protein refinements, was completed in three months. It is relatively inexpensive in computing time, in that although a fairly large number of cycles were calculated, each took little time (12–15 min early in refinement, 18–20 min later on, for *xyz* cycles; shorter times for *B*-refinement cycles). Most of the time taken for the refinement was occupied by manually refitting residues, locating solvent molecules, *etc.*

The early refinement cycles, with no manual intervention or rebuilding, produced a very rapid improvement in the structure. This reflects the fact that, given a reasonable starting model, most of the structure, particularly the main chain, refines quite automatically. Up to the time the first difference map was calculated ($R = 0{\cdot}29$), after seven cycles of *xyz* refinement, the average shift for main-chain atoms was 0·43 Å and for side-chain atoms 0·51 Å. From that point to the end of refinement the average shift for main-chain atoms was only 0·18 Å. It was during the latter period, however, that many side-chain conformations were manually corrected, and solvent molecules included. An example of the size of shifts that can be achieved automatically is shown by the side chain of Val 128 (Fig. 12*a*). The change in conformation shown occurred without any manual intervention during least-squares refinement, with shifts C^{α} 1·4, C^{β} 1·6, $C^{\gamma 1}$ 1·0, $C^{\gamma 2}$ 1·1 Å. It also illustrates one of the advantages of individual-atom refinement. If the side chain was refined as a group, a rotation of 180° would be required to effect the change shown in Fig. 12(*a*), but with unlinked atoms this is achieved simply by a large shift in C^{β} inverting the conformation. Similar shifts were observed for several Thr side chains (although $O^{\gamma 1}$ and $C^{\gamma 2}$ then had to be interchanged).

Even with a good starting model there are likely to be some errors too large to be corrected by automatic refinement procedures. A typical example (Ile 78) is shown in Fig. 12(*b*) and others have been discussed earlier (Glx 86, Fig. 8 and Ser 42, Fig. 12*c*). Thus an essential part of the refinement system is the ability to recognize such errors without intensive study of difference maps. (Our policy then was 'If in doubt, leave it out'.) In the earlier stages of refinement, high or inconsistent *B* values are the best guide (*e.g.* errors in peptide orientations were indicated by very high *B* values for the carbonyl oxygen atoms); later on, when most of the structure has refined, the observation of large shifts on refinement, reversed on regularization, can provide a second indicator.

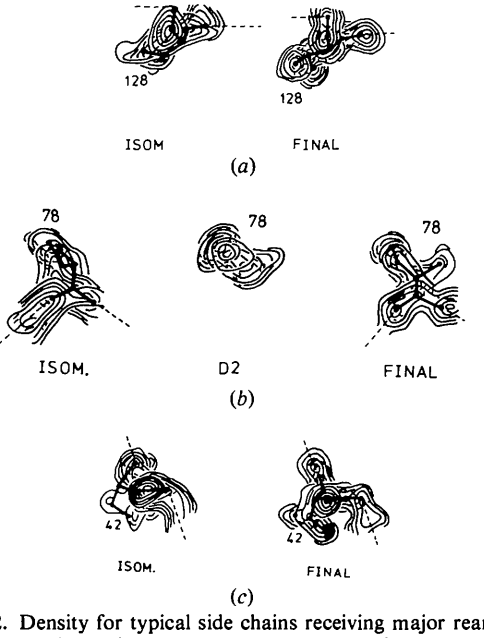


Fig. 12. Density for typical side chains receiving major rearrangements during refinement. In each case the fit to the original isomorphous density (left) is compared with the fit to a final difference map calculated after these residues had been temporarily omitted. In (*a*) Val 128 is seen. This side chain received large shifts (1·0–1·5 Å) during the automatic refinement, without any manual intervention. In (*b*) Ile 78 is shown. On refinement, the *B* values for $C^{\gamma 1}$ and $C^{\delta 1}$ increased rapidly, the side chain was then omitted, and the correct conformation became clear in a subsequent difference map (*D*2). In (*c*) Ser 42 is typical of a small number of side chains which have been given disordered conformations. Density in the isomorphous map was poor, and hard to fit. The residue refined automatically to the position shown on the right, but C^{β} tended to refine towards a second peak adjacent to it. This has been taken as a disordered position for O^{γ} (but note that an alternative explanation is that this could be a Thr residue, representing a sequence error).

A real strength of the least-squares refinement program is the ease with which individual *B* values can be refined. Not only are they frequently invaluable in pinpointing major errors in the structure, but when high-resolution data is available they give useful information about the mobility of parts of the structure. Since *B* values are highly correlated with the accuracy of atomic positions, however, they should not be refined too early as this may inhibit the refinement of atoms which are still misplaced (by flattening gradients).

The unreliability, early in refinement, of peaks which might be ascribed to solvent, suggests that solvent molecules should be introduced with caution. Although they form a significant part of the structure, their absence did not appear to affect the initial refinement adversely. Because the maximum size of shifts was limited (to two or three times the average), and the structure was regularized quite frequently, protein atoms were prevented from refining into the density due to solvent molecules.

We have not attempted to include bond-length constraints *etc.* in the least-squares refinement program [as in the methods developed by Konnert (1976)]. While this would presumably be of great benefit where the X-ray data were limited, there are some advantages to be gained from regularizing the structure outside the least-squares program (*i.e.* in using a two-stage process). In particular, large shifts on regularization sometimes were important in identifying structure which, despite moderate *B* values, contained errors too large to be corrected by automatic refinement. Bond-length and angle restraints may also prevent automatic refinement of the sort shown for Val 128.

It is true that the initial coordinate set in this case was of better than average accuracy, and this undoubtedly aided the early stages of refinement. Nevertheless, we strongly recommend the method for refinement of other protein structures, when data of resolution better than ~2·5 Å are available (and trials are in progress to assess its suitability when lower-resolution data only can be obtained). There seems no reason why refinement should not now be regarded as a routine part of protein-structure analyses.

We wish to thank Drs C. E. F. Rickard and P. E. Nixon of the University of Auckland, Dr K. L. Brown of the Chemistry Division, DSIR, Wellington, and Dr W. T. Robinson of the University of Canterbury, for help with the data collection. We are grateful for the use of the diffractometers in those centres.

One of us (ENB) wishes to thank the Massey University Council for granting and assisting a year's study leave, during which most of this work was done; the Chemistry Department of the University of York for provision of facilities; and the Royal Society and the Dame Lillian Penson Foundation for most generous additional financial assistance.

Most of all we wish to thank Drs R. Agarwal and N. Isaacs for their helpful advice and for providing us with manuscripts of their work prior to publication, and Dr G. Dodson for his constant enthusiasm and encouragement in this work.

References

ADMAN, E. T., SIEKER, L. C. & JENSEN, L. H. (1976). *J. Biol. Chem.* **251**, 3801–3806.

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

ANDERSON, C. M., MCDONALD, R. C. & STEITZ, T. A. (1978). *J. Mol. Biol.* **123**, 1–14.

ANDERSON, C. M., STENKAMP, R. E. & STEITZ, T. A. (1978). *J. Mol. Biol.* **123**, 15–34.

BAKER, E. N. (1974). *J. Mol. Biol.* **74**, 411–412.

BAKER, E. N. (1977). *J. Mol. Biol.* **115**, 263–277.

BODE, W. & SCHWAGER, P. (1975). *J. Mol. Biol.* **98**, 693–717.

CARNE, A. & MOORE, C. H. (1978). *Biochem. J.* **173**, 73–83.

CHAMBERS, J. L. & STROUD, R. M. (1977). *Acta Cryst.* B**33**, 1824–1837.

DEISENHOFER, J. & STEIGEMANN, W. (1975). *Acta Cryst.* B**31**, 238–250.

DIAMOND, R. (1966). *Acta Cryst.* **21**, 253–266.

DIAMOND, R. (1971). *Acta Cryst.* A**27**, 435–452.

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.

DRENTH, J., JANSONIUS, J. N., KOEKOEK, R. & WOLTHERS, B. G. (1971). *Adv. Protein Chem.* **25**, 79–115.

FREER, S. T., ALDEN, R. A., CARTER, C. W. & KRAUT, J. (1975). *J. Biol. Chem.* **250**, 46–54.

HUBER, R., KUKLA, D., BODE, W., SCHWAGER, P., BARTELS, K., DEISENHOFER, J. & STEIGEMANN, W. (1974). *J. Mol. Biol.* **89**, 73–101.

ISAACS, N. W. & AGARWAL, R. C. (1978). *Acta Cryst.* A**34**, 782–791.

KONNERT, J. H. (1976). *Acta Cryst.* A**32**, 614–617.

KRIEGER, M., CHAMBERS, J. L., CHRISTOPH, G. G., STROUD, R. M. & TRUS, B. L. (1974). *Acta Cryst.* A**30**, 740–748.

LUZZATTI, V. (1952). *Acta Cryst.* **5**, 802–810.

MOEWS, P. C. & KRETSINGER, R. H. (1975). *J. Mol. Biol.* **91**, 201–228.

NORTH, A. C. T., PHILLIPS, D. C. & MATHEWS, F. S. (1968). *Acta Cryst.* A**24**, 351–359.

RAMAKRISHNAN, C. & RAMACHANDRAN, G. N. (1965). *Biophys. J.* **5**, 909–933.

SHEFTER, E. (1970). *J. Chem. Soc. B*, pp. 903–906.

SUSSMANN, J. L., HOLBROOK, S. R., CHURCH, G. M. & KIM, S.-H. (1977). *Acta Cryst.* A**33**, 800–804.

SUSSMANN, J. L., HOLBROOK, S. R., WARRANT, R. W., CHURCH, G. M. & KIM, S.-H. (1978). *J. Mol. Biol.* **123**, 607–630.

TAKANO, T. (1977). *J. Mol. Biol.* **110**, 537–568.

TEN EYCK, L. F. (1973). *Acta Cryst.* A**29**, 183–191.

TEN EYCK, L. F. (1977). *Acta Cryst.* A**33**, 486–492.

WATENPAUGH, K. D., SIEKER, L. C., HERRIOTT, J. R. & JENSEN, L. H. (1973). *Acta Cryst.* B**29**, 943–956.