Acta Cryst. (1992). A**48**, 104–113

Fitting Electron Density by Systematic Search

BY RANDY J. READ

Department of Medical Microbiology and Infectious Diseases, University of Alberta, Edmonton, Alberta, Canada T6G 2H7

AND JOHN MOULT

Center for Advanced Research in Biotechnology, Maryland Biotechnology Institute, 9600 Gudelsky Drive, Rockville, MD 20850, USA

(*Received* 10 *December* 1990; *accepted* 2 *July* 1991)

Abstract

A systematic search approach to the automatic refinement of protein structures could reduce the need for manual intervention. In this approach, possible conformations for a segment of the polypeptide chain are generated systematically and the trial segments are scored for their agreement with the observed diffraction data. The sampling of conformational space is sufficiently exhaustive that reasonable conformations should be included. A number of score functions have been tested, including local electron-density correlations and global structure-factor agreements. The score functions vary in their predictive power as well as in their bias toward the conformation found in the current refined model, but the best score functions have reasonable predictive power. Related functions can be used to indicate which regions of the model fit poorly, reducing the need for manual inspection of models in electron density.

1. Introduction

In spite of continuing advances in automated refinement methods, macromolecular crystallographers still spend a large fraction of their time in the manual refitting of models to electron density. As refinement is only a means to an end, it would be much better to save the crystallographer's time for examination of the final refined model. In this paper, we describe a method to perform the refitting automatically. We show that this method gives useful results in a realistic test case.

Conventional automated refinement methods are myopic in the sense that only the effects of small movements of atoms are visible to the computer program at any time. The myopia results from algorithms relying on the gradients of the (non-linear) Fourier transform. Large conformational changes often cannot be achieved in small steps without encountering conformations that fit the observations poorly; if atoms must be moved from incorrect to correct density, they will be out of density at intermediate steps. Such situations create barriers to change and refinement can become trapped in local minima. On the other hand, the crystallographer examining an electron-density map recognizes patterns, not gradients, and applies in one step the large conformational changes that are necessary. Algorithms that avoid the reliance on gradient methods and small shifts should similarly be able to escape from local minima and take over more of the refinement task.

The technique of refinement by molecular dynamics (MD) (Brünger, Kuriyan & Karplus, 1987; Fujinaga, Gros & van Gunsteren, 1989) has made a substantial improvement from simple least-squares refinement. MD refinement has a larger radius of convergence, *i.e.* more progress is made before manual intervention is required. Atomic shift gradients are treated as forces; with the addition of thermal energy, unfavorable conformations are tolerated temporarily, allowing escape from at least some local minima. However, there are still limits to the scope of MD refinement and the goal of completely automated refinement has not yet been achieved. In particular, structures solved by molecular replacement techniques present problems in the regions of low homology. For example, errors in amino acid sequence alignment are not uncommon. To correct them requires the concerted shift of a number of atoms, many of which fit reasonably well into density. In the refinement of aspartate aminotransferase (Brünger, 1988), MD refinement was unable to shift residues out of density into which they had been modeled incorrectly. Thermal energy sufficient to move atoms out of incorrect density would also disrupt the correct regions of the model. These problems could be avoided by a method allowing large shifts of conformation that jump out of local minima, regardless of the height of the barriers.

We have chosen a systematic search approach to the refitting problem. A series of trial conformations is generated for a segment of the protein and each trial segment is evaluated. The systematic search algorithm used to generate the trial structures (Moult & James, 1986) will be described briefly in § 2. The

0108-7673/92/020104-10$03.00

major consideration remaining is the choice of the score function to evaluate the trial segment. A good score function should be as sensitive as possible to improvements in the model and it should preferably be efficient to evaluate. In § 3 we discuss a number of possible score functions and in § 4 we show the results of numerical tests.

2. The systematic search algorithm

The procedure for generating a set of conformations has been described in detail elsewhere (Moult & James, 1986). Here we summarize the algorithm and outline the adaptations necessary for a crystallographic application.

The objective is to sample the conformational space of the residues involved sufficiently finely that one can be assured that at least one conformation is generated with less than the maximum tolerable error. In applications where energy considerations are to be used in evaluating the correctness of conformations, all atoms of the residues must be built. An advantage of the crystallographic situation is that the problem may be subdivided into construction of main-chain conformations, selection of the best of these, and then construction of all possible side-chain conformations on that one main chain. On the other hand, crystallographic score functions are promiscuous: they can only detect whether *any* atom is close to a real atomic position, not whether the *correct* atom is near to such a position. This means that a fairly fine sampling of the conformational space is necessary.

The tests described here show that it is necessary to generate a conformation within a root-mean-square (r.m.s.) distance of approximately 0·7 Å from the correct one in order for the available score functions to perform successfully. Conformations are generated using main-chain building blocks with standard geometry (Sielecki *et al.*, 1979) and varying the main-chain φ and ψ dihedral angles. A set of 20 pairs of dihedral angle values for residues other than glycine and proline produces a sufficiently dense sampling (Table 1).

The values were selected manually by inspection of the distribution of observed values in the well ordered regions of 14 high-resolution well refined structures (Moult & James, 1986). They adequately sample the α, β, α/β bridge and left-handed α regions of the Ramachandran plot. A check against a set of 24 such proteins shows there are only 40 φ/ψ values more than 30° in φ or ψ from one of these building points in the regions with temperature factors lower than 20 Å^2 (Herzberg & Moult, 1991).

From this set of φ/ψ values, there are $3{\cdot}2\times10^6$ possible conformations for a five-residue loop, too large a number to be evaluated with the score functions. The number of conformations must therefore be restricted further by the use of restraints imposed by the rest of the structure. We have used three such rules:

Table 1. *φ/ψ values (°) used to generate main-chain conformations*

φ	ψ	φ	ψ	φ	ψ
−165	−165	−105	90	−75	120
−165	−135	−105	120	−75	150
		−105	150	−75	180
−135	90	−105	180		
−135	120			−60	−20
−135	150	−90	−45	−60	−50
−135	180	−90	−15		
		−90	15	60	30
−120	15				

(1) Conformations must come within specified limits of maintaining chain continuity across the region of chain considered, using the standard residue geometry.

(2) van der Waals overlaps with atoms of the rest of the protein must be less than some specified amount.

(3) No van der Waals overlaps are allowed within the loop.

The stretch of chain to be built is split into two pieces: for example, for a five-residue loop, two residues form the N-terminal portion and three residues the C-terminal portion. Each portion is then built independently onto the residue of known position on the abutting N- or C-terminal region of the chain (referred to as the root residue). Thus there are 400 possible conformations on the N-terminal side and 8000 on the C-terminal side. All 20 possible conformations of the first residue are built, then all 20 versions of the next residue are built onto the accepted conformations of the first residue, and so on. As each new conformation of a residue is generated, a check is made to determine if it will be possible to obtain chain closure and if there are unacceptable overlaps with the surrounding structure. Overlaps are checked using a previously generated 0·5 Å grid. At each grid position is stored the distance to the nearest atom in the surrounding structure. This arrangement increases the efficiency of the overlap checking.

When all acceptable conformations for the two portions have been generated, the distance from every end on the N-terminal portion to every end on the C-terminal portion is calculated to see if they are close enough for small adjustments of the main-chain geometry to allow loop closure. Two quantities are examined: the distance between the amide N atoms that should superimpose and the distance between the C^α atoms on either side of the join (to check alignment of the halves of the chain).

Each pair of matched N- and C-terminal portions of the loops is then subjected to 40 steps of energy minimization using bond length, bond angle and proper and improper dihedral energy terms, together

with a van der Waals energy term. No electrostatic term is included. The root residues at either end are included in the minimization, with atom positions restrained tightly to the experimental positions. This procedure distributes the strain of loop closure over the whole loop while maintaining the relationships to the polypeptide chain on either side. The energy minimization is carried out using the *GROMOS* package (van Gunsteren & Berendsen, 1987).

The resulting loop conformations are compared with each other. Where two conformations are found to be very similar, one is eliminated. For crystallographic work, similarity is defined in terms of the maximum difference in coordinates for any pair of equivalent atoms.

Other main-chain-conformation generating schemes are possible. This one has two principal strengths: the density of sampling of the conformational space is reasonably well defined for a given number of residues and φ/ψ points, unlike procedures that use random sampling (Shenkin, Yarmush, Fine, Wang & Levinthal, 1987), and the covalent geometry is allowed to relax from exact standard values to obtain loop closure, unlike procedures that use exact numerical solutions to obtain possible conformations (Gō & Scheraga, 1970) (see below). A rather different approach would be to use fragments of known structures to define the approximate course of the backbone (Jones & Thirup, 1986). That approach might work well for the main-chain atoms. However, the database is not yet large enough to provide an all-atom level of description of the polypeptide chain. Side chains would still have to be generated systematically.

Caution is required in all procedures that rely on filtering to reduce the number of conformations because of the possibility of eliminating the best ones through an overly stringent application of the rules. In the present algorithm, some van der Waals overlap with the surrounding structure is allowed to avoid this problem. In a crystallographic application, the acceptance of such atomic overlaps restricts the choice of score functions (see later).

Caution is also required in the use of ideal covalent geometry together with dihedral angles to generate conformations: if the loop 5 test region (described below) is built with the crystallographically determined dihedral angles, the resulting main-chain r.m.s. deviation from the crystallographic coordinates is 0·48 Å, with a maximum deviation of 0·94 Å, indicating that the distortions from ideal geometry play a significant role in the observed conformation. Since these residues have high apparent mobility in the crystal (B factors up to 35 Å^2), some of the distortion is the result of observing an average structure, rather than covalent strain in a single structure (Kuriyan, Petsko, Levy & Karplus, 1986). Nevertheless, it is the observed average structure we are trying to reproduce. Tests with the Gō & Scheraga (1970) procedure for closing three residue loops with ideal geometry on all three residue stretches of *Streptomyces griseus* trypsin show that approximately 10% of loops cannot be closed at all in this manner and the r.m.s. error for many more is poor. The energy-minimization step used in the present algorithm is designed to avoid this problem.

There are three significant differences in the procedure used here from the earlier version (Moult & James, 1986).

(1) More φ/ψ values have been used to ensure the finer sampling of conformational space needed for crystallographic score functions to be effective.

(2) Laxer criteria have been used for accepting loop ends as near enough to be annealed to a closed loop (see § 4*b*). It was found that some of the best conformations were rejected with the old values.

(3) A van der Waals energy term has been added to the energy function used for loop closure. This results in loop conformations with zero atomic overlap, desirable because atomic overlap is found to interfere with the crystallographic scoring functions.

3. Score functions

(*a*) *Local fit to electron density*

We begin with the idea that the computer will replace the crystallographer in the task of refitting. This suggests that the score function should, like the crystallographer, evaluate the local fit of the trial segment to electron density. A simple score is the sum of the electron-density values at the atomic positions. However, since both the model and the map have continuous electron-density distributions, this places unreasonable weight on the density at a single position. A continuous analog to the sum of densities is the electron-density product function, P.

$$P = \int_{\text{cell}} \rho_{\text{map}}\rho_{\text{seg}}\, \mathrm{d}V. \qquad (1)$$

ρ_{map} is an electron-density map that should show the true structure (several possible maps are discussed below) and ρ_{seg} is a map of the density of the trial structure of a segment. P can be evaluated efficiently in reciprocal space using Parseval's theorem.

$$P = 2/V \sum_{\text{hemisphere}} |F_{\text{map}}||F_{\text{seg}}| \cos(\alpha_{\text{map}} - \alpha_{\text{seg}}). \qquad (2)$$

A possible problem with the product function is that, using the systematic search algorithm, atoms of the trial segment may have significant overlap (see § 2). If a number of overlapped atoms fall into the same density, P increases unreasonably. This effect can be partially overcome by converting P to the coefficient of correlation between the two maps, where the denominator also increases when atoms overlap. In addition, the score is then on an absolute scale from

−1 to +1. All terms in the local map correlation coefficient can be computed in reciprocal space using Parseval's theorem, as for the product function. If the sum is taken over the unique set of reflections, each term should strictly be multiplied by the reflection multiplicity (the number of times symmetry equivalents of that reflection occur in the hemisphere). However, we approximate by simply taking the sum over the unique set of reflections. The mean density of the electron-density maps is subtracted within the correlation coefficient simply by omitting the reciprocal-lattice origin term (F_{000}) from the summations.

$$C = \frac{\sum |F_{map}||F_{seg}| \cos(\alpha_{map} - \alpha_{seg})}{[\sum |F_{map}|^2 \sum |F_{seg}|^2]^{1/2}}. \quad (3)$$

(b) *Choice of map coefficients*

The best choice of map coefficient (F_{map}) will maximize the discrimination between correct and incorrect versions of the trial segment. The usefulness of map weighting is well established. For instance, a figure-of-merit-weighted electron-density map (Blow & Crick, 1959) minimizes the r.m.s. deviation from the true map. We focus here on situations where a set of phases based on a partial model of the structures is available. In generating a map to assess trial structures we may use just the phases derived from the rest of the structure ($\mathbf{F}_{par}$) or we may include contributions from each trial structure for the region of interest ($\mathbf{F}_{p+s} = \mathbf{F}_{par} + \mathbf{F}_{seg}$). In the latter case, there will be model bias, tending to produce density corresponding to that trial structure. One might argue that it is best to omit the trial segment to reduce model bias. On the other hand, it has often been noted that incorrectly placed atoms fit poorly into density even when they are included in the phase calculation, while correctly placed atoms fit well, especially when they are included in the phase calculation. This argument implies that discrimination by the score functions could be improved by including the trial segments.

For a map that represents the total electron density in the unit cell, the following coefficients should reduce map errors and model bias simultaneously (Read, 1986).

$$F_{map} = (2m|F_{obs}| - D|F_{calc}|) \exp(i\alpha_{calc}). \quad (4)$$

In this expression, m is the figure of merit and D is a resolution-dependent factor that depends on the errors in the model used to compute F_{calc} (Luzzati, 1952; Read, 1990). The model could include or omit the trial segment. The factor D can be rationalized as follows. The best model (in a r.m.s. sense) of the true electron density is its expected value. To get the expected electron density from the atomic model, the atoms must be smeared out, or convoluted, by the overall uncertainty in their positions. Multiplication by the factor D achieves this convolution (Read, 1990).

A disadvantage is that a total electron-density map will give a correlation signal even if atoms in the trial segment fall into density that is already accounted for by other parts of the model. This can be circumvented by isolating the unaccounted density in a difference map. Then

$$F_{map} = (m|F_{obs}| - D|F_{par}|) \exp(i\alpha_{par}). \quad (5)$$

A discussion of these difference-map coefficients will be published elsewhere.

Phasing information from the trial segment can also be used for a difference map, but now we must use a vector difference coefficient, given by

$$F_{map} = m|F_{obs}| \exp(i\alpha_{p+s}) - D|F_{par}| \exp(i\alpha_{par}). \quad (6)$$

(c) *Global score functions*

The electron-density correlations described above limit the evaluation of a segment to its local density, a strategy that might be expected to optimize discrimination by the score function. However, a counter argument is that the general phase improvement from a correct segment will improve the accuracy of the density elsewhere in the map and thus that the discrimination might be better if agreement were evaluated globally.

A number of global score functions are used in crystallography to evaluate the agreement between model and observations. Of primary importance is the residual minimized during least-squares refinement, generally the weighted sum of squares of deviations of the amplitudes. If unit weights are used, the square root of the residual can be interpreted as a measure of the r.m.s. difference between the maps computed with coefficients $|F_{obs}| \exp(i\alpha_{p+s})$ and $\mathbf{F}_{p+s}$ (Wilson, 1976; Silva & Rossmann, 1985).

$$S = \sum (|F_{obs}| - |F_{p+s}|)^2. \quad (7)$$

On the other hand, the figure that is generally quoted for the quality of a model is the standard crystallographic R factor.

$$R = \sum ||F_{obs}| - |F_{p+s}|| / \sum |F_{obs}|. \quad (8)$$

However, for a relatively poor model such as an unrefined structure or a molecular replacement model, the correlation coefficient seems to be a better indicator (Fujinaga & Read, 1987; Cygler & Anderson, 1988). The coefficient of correlation between $|F_{obs}|$ and $|F_{p+s}|$ is defined as

$$C = \frac{\sum (|F_{obs}| - \overline{|F_{obs}|})(|F_{p+s}| - \overline{|F_{p+s}|})}{[\sum (|F_{obs}| - \overline{|F_{obs}|})^2 \sum (|F_{p+s}| - \overline{|F_{p+s}|})^2]^{1/2}}. \quad (9)$$

(*d*) *Refinement bias*

In examining maps computed using phases that omit a particular segment, one often has the impression that, in some sense, the phases 'remember' the missing segment and that there is a bias tending to reproduce its density. If the bias to reproduce the current structure were a significant factor, it could interfere with the systematic search approach. To avoid refinement bias, a few cycles of least-squares refinement excluding the segment (omit refinement) are often performed before computing a map (*e.g.* James, Sielecki, Brayer, Delbaere & Bauer, 1980). Brünger (1990) has recently studied this phenomenon more systematically and proposes that simulated annealing omit refinement is more effective at removing the bias.

Refinement bias can be rationalized as the result of compensating errors. If a segment of the structure is in error, refinement can distort the rest of the structure to provide a compensating error that reduces disagreement in the amplitudes. This is possible only because the ratio of observations to parameters is low. As a result of the error compensation, the influence of the incorrect segment on the rest of the structure remains even after it is removed and the electron density that will satisfy the amplitude constraints tends to resemble the density of the incorrect segment. The improvement in the agreement of the amplitudes must be achieved at the expense of phase accuracy. In fact, amplitudes agree better after refinement than one would expect from the phase error (Lunin & Urzhumtsev, 1984; Read, 1986).

Refinement bias could potentially complicate efforts to fit density automatically, causing loops that resemble the current conformation to be scored anomalously high. To check for this possibility, we have performed tests both before and after omit refinement.

4. Numerical tests

(*a*) *Test structure*

For a test of the automatic refitting algorithm and of the various possible score functions, we chose a realistic test case: we have fitted two of the surface loops of *Streptomyces griseus* trypsin (SGT), a serine proteinase related to bovine trypsin (BT). The structure of SGT was solved primarily by molecular replacement (Read & James, 1988), using the known structure of BT (Chambers & Stroud, 1979) as one of the models. After seven cycles of least-squares refinement with *PROLSQ* (Hendrickson & Konnert, 1980), the model was incomplete, having only 204 of 223 amino acid residues, and in some regions it resembled BT more than SGT. The conventional crystallographic R factor [6·0–2·8 Å data having $I > 3\sigma(I)$] was 0·425. Tests at this stage of refinement showed that the phases were too poor to provide a useful signal (see results below). However, at cycle 15, when 200 residues were included in a model giving an R factor of 0·351, a useful signal was obtained. We have used this model for most of our tests. Although manual intervention was required to reach this point in the original least-squares refinement, a MD refinement would be expected to get at least that far without rebuilding. To investigate the effect of refinement bias (§ 3*d*), five cycles of least-squares refinement were carried out on the cycle 15 model, omitting loop 5 (see below), to give what was termed the cycle 20a model.

Several loops that were modeled incorrectly early in the refinement have been used previously in tests of the systematic search algorithm (Moult & James, 1986). Only the main-chain component of the model is built here. Extension to the inclusion of side-chain sets would proceed by building all possible side-chain conformations onto the best main chain(s) selected, using the procedure described by Moult & James (1986).

(*b*) *Generation of trial conformations*

Conformations have been generated for the six-residue segment Gln110-Pro-Ile-Asn-Gln-Pro119 [chymotrypsinogen numbering of SGT (Read & James, 1988)], referred to as loop 4, and for the five-residue segment Asp203-Asn-Ala-Asp-Glu206 (loop 5).

Loop 5 is particularly interesting because it is possible to construct a large number of different main-chain conformations that satisfy the end constraints. In some respects we expect this exposed surface loop to be a particularly challenging example because the final electron density suggests that it is one of the least well ordered regions of the structure (Read & James, 1988); the mean B factor for main-chain atoms is 29·9 Å^2, with a maximum of 38·7 Å^2. Main-chain conformations were built using the procedure described in § 2. 380 conformations out of the possible 400 for the N-terminal two residues were accepted. The others were rejected because of excessive clashes with the surrounding structure. On the C-terminal side, 358 conformations of the first two residues were accepted with 42 clash rejects. The third residue was built onto these 358 new roots and 1457 conformations were accepted, with a further 533 clash rejects and 5170 rejected because they could not result in loop closure (some conformations are rejected by both criteria).

Comparison of the end positions of the 358 N-terminal portions with the 1457 C-terminal portions resulted in the selection of 8995 combinations with acceptable geometry for input to energy minimization. The geometry was considered acceptable if the common amide N atoms were within 1·8 Å of each

Table 2. *Evaluation of various score functions*

Score function	R.m.s. error in C_{final} [equation (11), see text] for top 25 scores: Loop 4, cycle 15	Loop 5, cycle 7	Loop 5, cycle 15	Loop 5, cycle 20a
Map correlations against total density				
$F_{map}=(2\|F_{obs}\|-\|F_{par}\|)\exp(i\alpha_{par})$	0·0066	0·0203	0·0107	0·0093
$F_{map}=(2\|F_{obs}\|-\|F_{p+s}\|)\exp(i\alpha_{p+s})$	0·0018	0·0259	0·0088	0·0136
$F_{map}=(2m\|F_{obs}\|-D\|F_{par}\|)\exp(i\alpha_{par})$	0·0023	0·0222	0·0085	0·0093
$F_{map}=(2m\|F_{obs}\|-D\|F_{p+s}\|)\exp(i\alpha_{p+s})$	0·0029	0·0252	0·0105	0·0108
Map correlations against difference density				
$F_{map}=(\|F_{obs}\|-\|F_{par}\|)\exp(i\alpha_{par})$	0·0049	0·0171	0·0133	0·0060
$F_{map}=(m\|F_{obs}\|-D\|F_{par}\|)\exp(i\alpha_{par})$	0·0035	0·0127	0·0068	0·0061
$F_{map}=m\|F_{obs}\|\exp(i\alpha_{p+s})-D\|F_{par}\|\exp(i\alpha_{par})$	0·0018	0·0127	0·0094	0·0092
Global scores				
$R=\sum\|\|F_{obs}\|-\|F_{p+s}\|\|/\sum\|F_{obs}\|$	0·0012	0·0194	0·0171	0·0097
Correlation of $\|F_{obs}\|$ with $\|F_{p+s}\|$	0·0033	0·0201	0·0072	0·0084
R.m.s. $(\|F_{obs}\|-\|F_{p+s}\|)$	0·0015	0·0286	0·0116	0·0110

other and the C^{α} atoms on either side of the bridge were between 2 and 6 Å apart.

After 40 steps of steepest descents energy minimization, the resulting complete loop conformations were compared with each other and conformations were rejected if all atoms were closer than 0·5 Å to their equivalents. 1690 conformations were finally accepted for examination using the crystallographic score functions. The best of these has a r.m.s. difference from the final X-ray structure of 0·60 Å for all non-H atoms of the loop 5 residues, the worst a r.m.s. of 3·15 Å.

Loop 4 is well ordered in the final structure (main chain mean $B=16{\cdot}2$ Å^2, maximum $B=25{\cdot}7$ Å^2) and forms an irregular extended conformation on the surface of the molecule. The presence of two proline residues, together with a relatively long end-to-end distance, means that fewer main-chain conformations are possible than for loop 5. The same procedure as for loop 5 was used to construct a set of main-chain conformations. 373 conformations were accepted for three residues on the N-terminal side of the segment and 182 for the three-residue C-terminal section. 1884 combinations were close enough for annealing into complete loops and after minimization and elimination of similar conformations, 392 were finally considered unique. Of these, the most accurate has a r.m.s. difference of 0·73 Å from the final structure and the worst a r.m.s. of 2·54 Å.

(*c*) *Evaluation of score functions*

To evaluate the score functions, we examine how well they predict the agreement of each trial segment with the final structure. The ideal agreement index is the r.m.s. distance of the atoms from their final positions. We shall use this measure as the definitive test of the method. However, as pointed out above, no score function can distinguish between an atom falling in its own density or in that of a different atom. Therefore, we evaluate the relative merits of score functions with an agreement index based on how well a trial segment accounts for the density missing from the current model. A map correlation is used for this.

For this purpose we need a map that represents the missing density. The density map that best represents the true structure is the Fourier transform of $m_{final}|F_{obs}|\exp(i\alpha_{final})$. The map showing our current knowledge of the structure, excluding the trial segment, is the Fourier transform of $D|F_{par}|\exp(i\alpha_{par})$. (As discussed above, the factor D allows for the effect of overall coordinate error.) So the difference map showing the density missing from the current model is the Fourier transform of

$$F_{diff}=m_{final}|F_{obs}|\exp(i\alpha_{final})-D|F_{par}|\exp(i\alpha_{par}). \quad (10)$$

The correlation coefficient between this difference map and the electron density of the trial segment provides our standard

$$C_{final}=\frac{\sum|F_{diff}||F_{seg}|\cos(\alpha_{diff}-\alpha_{seg})}{[\sum|F_{diff}|^2\sum|F_{seg}|^2]^{1/2}}. \quad (11)$$

(Note that C_{final}, in practice, does not change much when the weighting factors m_{final} and D are omitted.)

Scatter plots comparing the score functions with the final agreement indices, such as those in Fig. 1, are useful for a qualitative evaluation. An overall quantitative evaluation could be given by a correlation coefficient. However, what matters is that the few best trial segments be among the top scoring segments. We have evaluated this criterion as follows. The list of trial segments is sorted according to the value of C_{final}, with the best choice at the top of the list. A second sort of the list is made according to one of the score functions. The r.m.s. difference between the values of C_{final} on the first n lines of the two sorted lists will be a measure of how well their orders agree for the best n scores. If the relationship between C_{final} and the score function is monotonic, the lists will be in the same order and the r.m.s. difference will be zero. Table 2 gives this evaluation of various score functions for the test cases, with $n=25$.

From the scatter plots in Fig. 1 and the data in Table 2, we see that there is good discrimination by a number of the score functions. The trial models with high scores generally account well for the missing density, having atoms near the positions of atoms in the final model. Fig. 2 shows stereo views of selected models. What is somewhat surprising, and encouraging, is that even though the final model of loop 5 fits

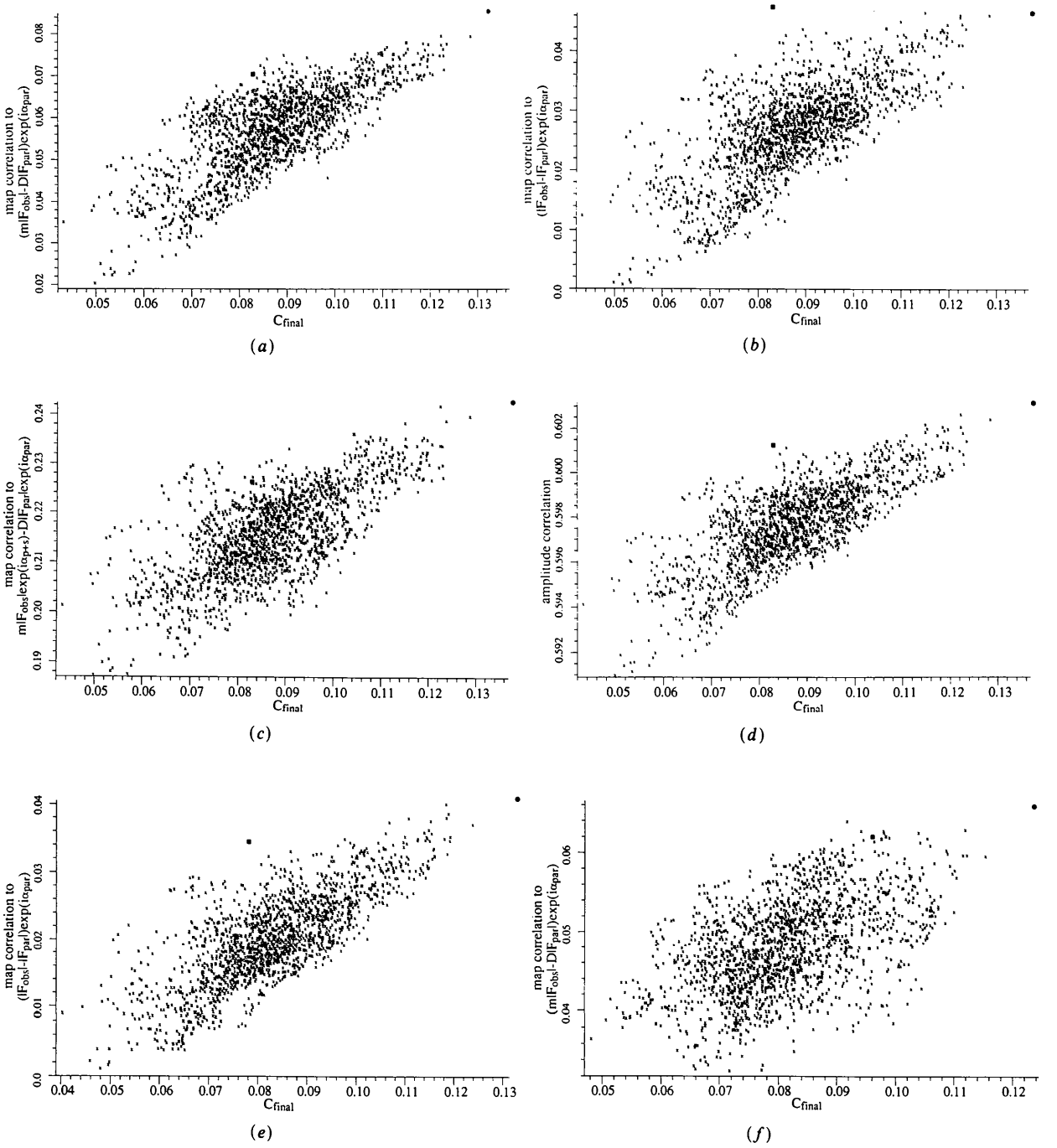


Fig. 1. Scatter plots comparing selected score functions with C_{final} [equation (11), see text] for the trial segments of loop 5 (crosses), the final model (dot) and the current model (square). The tested score functions are: (*a*) map correlation using $F_{\text{map}} = (m|F_{\text{obs}}| - D|F_{\text{par}}|)\exp(i\alpha_{\text{par}})$ at cycle 15; (*b*) map correlation using $F_{\text{map}} = (|F_{\text{obs}}| - |F_{\text{par}}|)\exp(i\alpha_{\text{par}})$ at cycle 15; (*c*) map correlation using $F_{\text{map}} = m|F_{\text{obs}}|\exp(i\alpha_{p+s}) - D|F_{\text{par}}|\exp(i\alpha_{\text{par}})$ at cycle 15; (*d*) correlation between $|F_{\text{obs}}|$ and $|F_{p+s}|$ at cycle 15; (*e*) map correlation using $F_{\text{map}} = (|F_{\text{obs}}| - |F_{\text{par}}|)\exp(i\alpha_{\text{par}})$ at cycle 20a; (*f*) map correlation using $F_{\text{map}} = (m|F_{\text{obs}}| - D|F_{\text{par}}|)\exp(i\alpha_{\text{par}})$ at cycle 7.

poorly into the cycle 15 density (Fig. 2*a*), so that manual refitting would be difficult, the score functions are giving useful results.

The following points arise from a comparison of the possible score functions:

(1) The three global scores considered are of similar efficacy. The amplitude correlation would be preferred because of its insensitivity to scaling errors.

(2) Generally, the map correlations using difference density are better than those using the total density.

(3) As expected, the use of weight factors (m and D) in the map coefficients improves the discrimination by the density correlation scores, at least before any omit refinement has been carried out. This can be seen, for instance, from a comparison of the results at cycle 15 with weighted and unweighted difference-map coefficients [compare Figs. 1(*a*) and (*b*)].

(4) Addition of the trial segment to the phase calculation does not appear to improve the discrimination of correct conformations from incorrect ones. The map correlations are higher when the segment is included, but for loop 5 there is also more scatter [compare Figs. 1(*a*) and (*c*)].

(5) Omission of the current loop coordinates for several cycles of refinement only gives a significant improvement for the map correlations using unweighted Fourier coefficients. This is particularly noticeable for the difference map [compare Figs. 1(*b*) and (*e*)].

(6) Although omit refinement improves the results with unweighted difference maps, the results are no better than those achieved without refinement but using weights [compare Figs. 1(*a*) and (*e*)].

(7) The scores are very sensitive to the quality of the phases: little useful information can be obtained for loop 5 at cycle 7 of refinement [compare Figs. 1(*a*) and (*f*)].

One of the best score functions is the map correlation which is computed using $F_{map} = (m|F_{obs}| - D|F_{par}|) \exp(i\alpha_{par})$. This score function is one of the cheapest to compute, it is basically insensitive to refinement bias and it is completely insensitive to misscaling of the data.

Having selected a scoring function that agrees well with the theoretical optimum, we can now address the question of how well it selects trial segments with the lowest r.m.s. deviation from the true structure. Fig. 3 shows there is a strong signal. The segments with the lowest r.m.s. deviation are found among the very highest scores. Note that the final structure scores best in both cases, indicating that the scatter among the highest-scoring trial segments is primarily due to the approximate structure rather than the poor quality of the phases. Thus, refinement of the top scoring trials would be expected to lead to a single highest scoring very low r.m.s. structure.

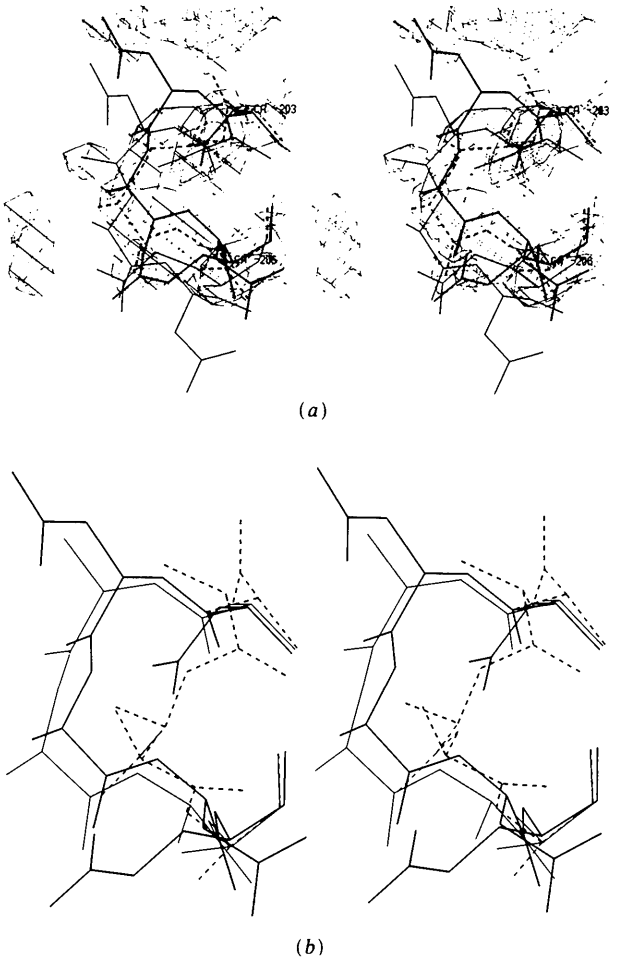
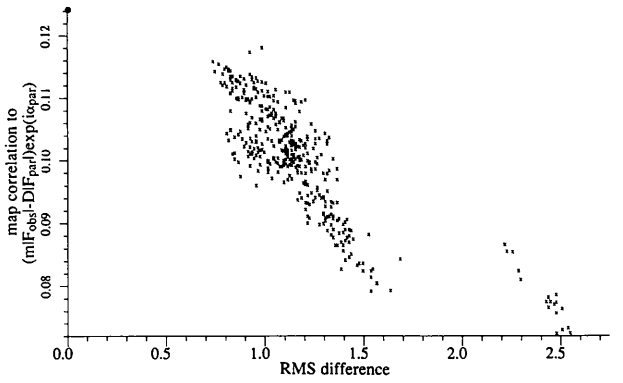


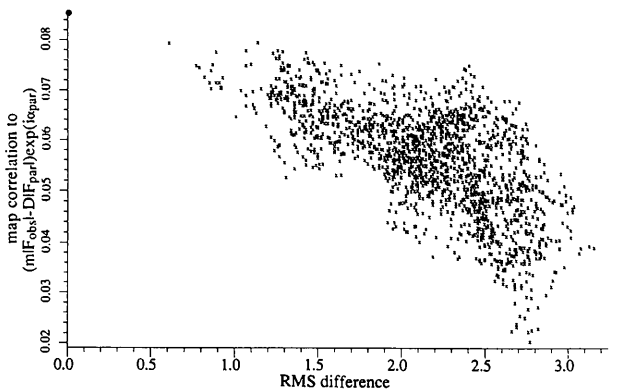
Fig. 2. Stereo views of loop 5 in SGT. The final conformation is in thick lines. The score used to select trial conformations is the map correlation with $F_{map} = (m|F_{obs}| - D|F_{par}|) \exp(i\alpha_{par})$. (*a*) Electron density is from a map computed at cycle 15 of least-squares refinement with coefficients $(2m|F_{obs}| - D|F_{par}|) \exp(i\alpha_{par})$. In thin lines is the conformation at cycle 15. The conformation with the highest score is shown in dashed lines. Note that the residue that agrees most poorly with the density, Asn204, has a mean main-chain B factor of 35·0 Å^2. (*b*) The conformation with the lowest r.m.s. deviation from the final structure (and the second-highest score) is shown in thin lines. The conformation with the lowest score is in dashed lines.

(*d*) *Automatic identification of poorly fit regions*

To do automatic, or manual, refitting, one must first identify the regions of the structure that fit poorly. A score related to the local scores discussed above should be useful. One of the best local scores is the map correlation defined with $F_{map} = (m|F_{obs}| - D|F_{par}|) \exp(i\alpha_{par})$. However, a score that will be used to compare regions of the structure should not be affected by differences in the types or numbers of atoms in the different segments. But it can be seen from (3) that the map correlation for a perfect model would vary with the number of atoms. A better score for comparing different regions would



(*a*)



(*b*)

Fig. 3. Scatter plots comparing the r.m.s. deviation from the final structure with the map correlation using $F_{map} = (m|F_{obs}| - D|F_{par}|) \exp(i\alpha_{par})$ at cycle 15 for (*a*) loop 4 and (*b*) loop 5. Crosses correspond to the trial segments and a dot to the final model.

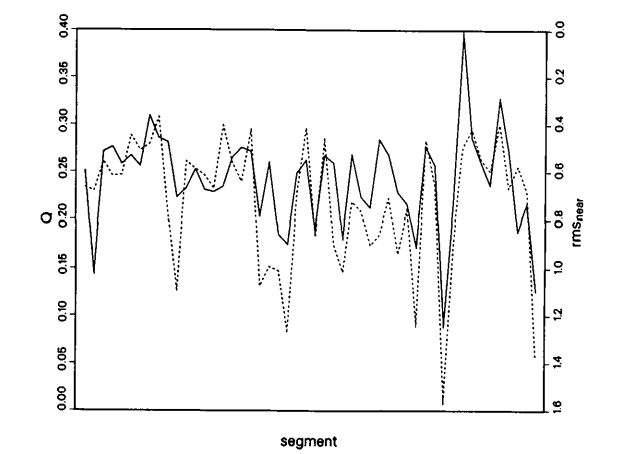


Fig. 4. Comparison of the quality-of-fit index Q (solid line), defined in equation (12), with r.m.s.$_{near}$ (dashed line) for four-residue segments of SGT at cycle 15 of least-squares refinement.

be the electron-density product function, normalized by the mean square density of the test region. This could be evaluated in reciprocal space as the following:

$$Q = \frac{\sum |F_{map}||F_{seg}| \cos(\alpha_{map} - \alpha_{seg})}{\sum |F_{seg}|^2}. \qquad (12)$$

Q would have a value of 1 for a perfect model (if the atoms were resolved and the scale were correct), independent of the number and type of atoms in the test segment.

To test this quality-of-fit score, the model of SGT at cycle 15 of least-squares refinement was divided into four-residue segments, for which Q was evaluated. Q was compared to r.m.s.$_{near}$, the r.m.s. distance to the nearest atom in the final model. Fig. 4 compares the variation in Q and r.m.s.$_{near}$ along the sequence and demonstrates that Q is a useful indicator of the quality of fit. The coefficient of correlation between Q and r.m.s.$_{near}$ is $-0{\cdot}726$. In comparison, when the corresponding map correlation is used as a quality-of-fit score, the corrrelation to r.m.s.$_{near}$ is only $-0{\cdot}650$. We have not attempted to compare Q with the real-space R factor, defined by Brändén & Jones (1990).

(*e*) *CPU requirements*

Our objective was to test the feasibility of automatic electron-density fitting, not to develop an efficient and user-friendly program at this stage. There is a great deal of extra computing overhead in the current implementation, particularly in the scoring program, which computes multiple alternative score functions and test statistics. The CPU timings, therefore, give extreme upper bounds on the actual requirements.

The generation and energy minimization of the loop conformations are reasonably efficient, taking less than 2 s on an IRIS 4D/25 to produce one of the loop 5 trial conformations. On the other hand, the scoring routine takes 30 s on one processor of an IRIS 4D/120 to evaluate each loop 5 conformation.

The programs to generate and evaluate trial conformations are available from the authors. *GROMOS*, which was used for energy minimization, is available from van Gunsteren & Berendsen (1987).

5. Concluding remarks

For the two segments of SGT examined, the systematic search procedure generates conformations close to the correct ones, and the score functions identify these. Although the lowest r.m.s. segments are not selected unerringly, they are always very near the top of the list. This means that, in practice, local refinement of the top scoring segments would be required before a final selection could be made. Only main-chain searches have been included in these trials

but the extension to include side chains is straightforward (Moult & James, 1986).

We envisage a procedure in which poor regions of the current model are automatically identified then replaced by the systematic search procedure. Following a round of such replacement, conventional refinement would be carried out. This would be followed by further rounds of systematic search and refinement until no further rebuilding occurs. The occasional failure to select a conformation within the convergence radius of the refinement would not be a serious obstacle since the general improvement of the phases would increase the power of the method in succeeding rounds. Thus, systematic search offers an alternative approach to extending the convergence radius of refinement, beyond even that offered by MD refinement.

RJR is an Alberta Heritage Foundation for Medical Research Scholar. This work was supported in part by NIH grant R01 GM41034 to JM and MRC (Canada) grant MT 11000 to RJR.

References

BLOW, D. M. & CRICK, F. H. C. (1959). *Acta Cryst.* **12**, 794-802.
BRÄNDÉN, C.-I. & JONES, T. A. (1990). *Nature (London)*, **343**, 687-689.
BRÜNGER, A. T. (1988). *J. Mol. Biol.* **203**, 803-816.
BRÜNGER, A. T. (1990). In Lecture Notes for the Crystallographic Computing School in Bischenberg, pp. 359-372.
BRÜNGER, A. T., KURIYAN, J. & KARPLUS, M. (1987). *Science*, **235**, 458-460.
CHAMBERS, J. L. & STROUD, R. M. (1979). *Acta Cryst.* B**35**, 1861-1874.
CYGLER, M. & ANDERSON, W. F. (1988). *Acta Cryst.* A**44**, 300-308.
FUJINAGA, M., GROS, P. & VAN GUNSTEREN, W. F. (1989). *J. Appl. Cryst.* **22**, 1-8.
FUJINAGA, M. & READ, R. J. (1987). *J. Appl. Cryst.* **20**, 517-521.
GŌ, N. & SCHERAGA, H. A. (1970). *Macromolecules*, **3**, 178-187.
GUNSTEREN, W. F. VAN & BERENDSEN, H. J. C. (1987). *GROMOS*: Groningen Molecular Simulation Library Manual. BIOMOS BV, Nijenborgh 16, Groningen, The Netherlands.
HENDRICKSON, W. A. & KONNERT, J. H. (1980). In *Biomolecular Structure, Function, Conformation and Evolution*, edited by R. SRINIVASAN, Vol. I, pp. 43-57. Oxford: Pergamon Press.
HERZBERG, O. & MOULT, J. (1991). *Proteins*, **11**, 223-229.
JAMES, M. N. G., SIELECKI, A. R., BRAYER, G. D., DELBAERE, L. T. J. & BAUER, C.-A. (1980). *J. Mol. Biol.* **144**, 43-88.
JONES, T. A. & THIRUP, S. (1986). *EMBO J.* **5**, 823-826.
KURIYAN, J., PETSKO, G. A., LEVY, R. M. & KARPLUS, M. (1986). *J. Mol. Biol.* **190**, 227-254.
LUNIN, V. Y. & URZHUMTSEV, A. G. (1984). *Acta Cryst.* A**40**, 269-277.
LUZZATI, V. (1952). *Acta Cryst.* **5**, 802-810.
MOULT, J. & JAMES, M. N. G. (1986). *Proteins*, **1**, 146-163.
READ, R. J. (1986). *Acta Cryst.* A**42**, 140-149.
READ, R. J. (1990). *Acta Cryst.* A**46**, 900-912.
READ, R. J. & JAMES, M. N. G. (1988). *J. Mol. Biol.* **200**, 523-551.
SHENKIN, P. S., YARMUSH, D. L., FINE, R. M., WANG, H. & LEVINTHAL, C. (1987). *Biopolymers*, **26**, 2053-2085.
SIELECKI, A. R., HENDRICKSON, W. A., BROUGHTON, C. G., DELBAERE, L. T. J., BRAYER, G. D. & JAMES, M. N. G. (1979). *J. Mol. Biol.* **134**, 781-804.
SILVA, A. M. & ROSSMANN, M. G. (1985). *Acta Cryst.* B**41**, 147-157.
WILSON, A. J. C. (1976). *Acta Cryst.* A**32**, 781-783.

Acta Cryst. (1992). A**48**, 113-120

Molecular Reorientation in an Electric Field as Studied by Single-Crystal X-ray Diffraction

BY H. GRAAFSMA,* A. PATURLE,† L. WU, H.-S. SHEU,‡ J. MAJEWSKI,§ G. POORTHUIS¶ AND P. COPPENS*

Department of Chemistry, State University of New York at Buffalo, Buffalo, NY 14214, *USA*

(*Received* 5 *April* 1991; *accepted* 17 *July* 1991)

Abstract

The molecular reorientation induced by an external electric field has been determined for the first time in order to obtain a microscopic understanding of the interaction between crystals and electric fields. Changes in scattering intensity are found when an electric field is applied parallel to the polar axis of a non-linear optical crystal of 2-methyl-4-nitroaniline (MNA), which has large piezoelectric constants [Paturle, Graafsma, Sheu, Coppens & Becker (1991). *Phys. Rev. B*, **43**, 14683-14691]. The effect has been analyzed in terms of a change in cell parameters, a molecular rotation of $0.45(5)\times10^{-2}$° about an axis nearly parallel to the electric field and a molecular translation of $0.19(3)\times10^{-3}$ Å along the *b* axis. Since

* Authors to whom correspondence should be addressed.
† Current address: 60 Boulevard Aristide Briand, 63000 Clermont-Ferrand, France.
‡ Permanent address: Synchrotron Radiation Research Center, 8th floor, No. 6, Roosevelt Road, Sec. 1, Taipei 10757, Taiwan.
§ Associated with: Department of Structural Chemistry, Weizmann Institute of Science, Rehovot 76100, Israel.
¶ Permanent address: Bentheimerstraat 75, 7587 NG, de Lutte, The Netherlands.

0108-7673/92/020113-08$03.00