J. Mol. Biol. (1986) **190**, 227–254

Effect of Anisotropy and Anharmonicity on Protein Crystallographic Refinement

An Evaluation by Molecular Dynamics

John Kuriyan[1,2], Gregory A. Petsko[1], Ronald M. Levy[3] and Martin Karplus[2]†

[1]*Department of Chemistry*
M.I.T., Cambridge, MA 02139, U.S.A.
[2]*Department of Chemistry*
Harvard University, Cambridge, MA 02138, U.S.A.
[3]*Department of Chemistry*
Rutgers University, New Brunswick, NJ 80903, U.S.A.

(Received 15 August 1985, and in revised form 17 December 1985)

Molecular dynamics simulations are employed to determine the errors introduced by anharmonicity and anisotropy in the structure and temperature factors obtained for proteins by refinement of X-ray diffraction data. Simulations (25 ps and 300 ps) of met-myoglobin are used to generate time-averaged diffraction data at 1·5 Å resolution. The crystallographic restrained-parameter least-squares refinement program PROLSQ is used to refine models against these simulated data. The resulting atomic positions and isotropic temperature factors are compared with the average structure and fluctuations calculated directly from the simulations. It is found that significant errors in the atomic positions and fluctuations are introduced by the refinement, and that the errors increase with the magnitude of the atomic fluctuations. Of particular interest is the fact that the refinement generally underestimates the atomic motions. Moreover, while the actual fluctuations go up to a mean-square value of about 5 Å^2, the X-ray results never go above approximately 2 Å^2. This systematic deviation in the motional parameters appears to be due to the use of a single-site isotropic model for the atomic fluctuations. Many atoms have multiple peaks in their probability distribution functions. For some atoms, the multiple peaks are seen in difference electron density maps and it is possible to include these in the refinement as disordered residues. However, for most atoms the refinement fits only one peak and neglects the rest, leading to the observed errors in position and temperature factor. The use of strict stereochemical restraints is inconsistent with the average dynamical structure; nevertheless, refinement with tight restraints results in structures that are comparable to those obtained with loose restraints and better than those obtained with no restraints. The results support the use of tight stereochemical restraints, but indicate that restraints on the variation of temperature factors are too restrictive.

1. Introduction

It is now recognized from a variety of experimental and theoretical studies that significant atomic motions occur in macromolecules of biological interest. Information concerning both the magnitudes and the time-scales of the motions are available (Karplus & McCammon, 1981, 1983). At room temperature, thermal atomic displacements are in the range of 0·2 to 2·0 Å and vary significantly for different regions of the protein; their time-scale is from 0·1 to 50 picoseconds (with the longer time-scales generally associated with the large amplitudes). The Debye–Waller (temperature) factors evaluated in X-ray crystallographic refinements of protein structures are an important source of experimental data concerning the magnitudes of the fluctuations (Petsko & Ringe, 1984); this is based on the identification of the temperature factors with the mean-square fluctuations of individual atoms. With the assumption of isotropic

† Author to whom all correspondence should be addressed.

0022-2836/86/140227-28 $03.00/0

and harmonic motion, temperature factors for all the non-hydrogen atoms have been determined for many proteins and some examples are given by Artymiuk *et al.* (1979), Frauenfelder *et al.* (1979), Watenpaugh *et al.* (1980), Takano & Dickerson (1981) and Sheriff *et al.* (1985).

It is clear, however, from molecular dynamics simulations that the atomic motions are highly anisotropic (Karplus & McCammon, 1981, 1983; Northrup *et al.*, 1981; van Gunsteren & Karplus, 1982*a*; Ichiye & Karplus, unpublished results) and, at least for some of the atoms, strongly anharmonic (Levy *et al.*, 1985; Mao *et al.*, 1982; van Gunsteren & Karplus, 1982*a,b*; Ichiye & Karplus, unpublished results). Since neither of these deviations from the simple model is included in most X-ray determinations of the structures of macromolecules, it is of interest to evaluate the errors introduced in the refinement process by their neglect. Such errors can involve the positions of atoms as well as their motional properties.

An evaluation of the errors is particularly important now that improved data sets can be obtained for macromolecules and more attention is being focused on deducing the motional properties by refinement of these data. With the advent of new techniques, such as the use of area detectors (Xuong *et al.*, 1978), low-temperature crystallography (Hartmann *et al.*, 1982) and intense X-ray sources available from synchrotrons (Wilson *et al.*, 1983), significant improvements in the quality of protein diffraction data are expected. For several small proteins, such as bovine pancreatic trypsin inhibitor, rubredoxin and crambin, the reflections have been measured out to 1·2 to 1·0 Å (Wlodawer *et al.*, 1984; Watenpaugh *et al.*, 1980; Teeter & Hendrickson, 1979) and it is possible also to collect high-resolution data (i.e. 1·2 to 1·0 Å) for larger proteins such as ribonuclease, lysozyme and myoglobin at low temperature (R. F. Tilton & J. Dewan, personal communication). This will make it possible to probe more deeply into the nature of protein motions and their effects on the X-ray refinement procedure. For a few proteins, anisotropic harmonic temperature factors have been introduced, resulting in six thermal parameters per atom (e.g. see Watenpaugh *et al.*, 1980). Anharmonic corrections have not been used for proteins, although they have been employed in small molecule refinements (Zucker & Schulz, 1982).

A direct experimental estimate of the errors resulting from the assumption of isotropic, harmonic temperature factors is difficult because sufficient data are not available for protein crystals. Moreover, any data set includes other errors, which would obscure the analysis, and the specific correlation of temperature factors and motion is complicated by the need to account for static disorder in the crystal. As an alternative to an experimental analysis of the errors in refinement procedures for proteins, we describe here a theoretical approach. The basic idea is to generate X-ray data from the results of a molecular dynamics simulation of a protein and to use these data to obtain a refined structure by standard methods. The error in the analysis is determined by comparing the results obtained from the refinement procedure with the known average structure and the mean-square fluctuations of the original simulation. This type of comparison, in which no real experimental results are used, avoids problems due to inaccuracies in the measured data (exact calculated intensities are used), crystal disorder (there is none in the model), and due to approximations in the simulation (the simulation gives exact results for this case). The only question about such a comparison is whether the atomic motions found in the simulation are a meaningful representation of those occurring in proteins. A variety of comparisons (Karplus & McCammon, 1981, 1983; Levy & Keepers, 1986) suggests that molecular dynamics simulations provide a reasonable picture of the motions, in spite of the errors in the potentials, neglect of the crystal environment and the finite time classical trajectories used to obtain the results. However, as already stated, these inaccuracies do not affect the exactness of the computer "experiments" and their interpretation given in this paper. This strategy is similar to that used previously to analyse transition state theory models for reactions (Morokuma & Karplus, 1971) and nuclear magnetic resonance relaxation models (Levy *et al.*, 1981).

A 25 picosecond molecular dynamics trajectory for myoglobin is used to carry out the test of the refinement procedure outlined above; the tests were also done using a 300 picosecond trajectory of myoglobin, but the results of the shorter simulation will be the focus of most of the discussion. The average structure and the mean-square fluctuations from that structure are calculated directly from the trajectory. To obtain the average electron density, appropriate atomic electron distributions are used for the individual atoms in each co-ordinate set the trajectory and averaged. Given the symmetry, unit cell dimensions and position of the myoglobin molecule in the cell, average structure factors are calculated as the Fourier transform of the averaged electron densities (van Gunsteren *et al.*, 1983). The resulting intensities at the Bragg reciprocal lattice points are used as input data for the widely applied crystallographic program PROLSQ (Konnert & Hendrickson, 1980). The time-averaged atomic positions obtained from the simulation and a uniform temperature factor provide the initial model structure. The positions and an isotropic, harmonic temperature factor for each atom are then refined iteratively against the computer-generated intensities in the standard way. PROLSQ is a restrained-parameter, least-squares refinement program, and the refinements are done with tight, loose and no restraints on the parameters.

Differences between the refined results for the average atomic positions and their mean-square fluctuations and those obtained from the molecular dynamics trajectory are due to errors introduced by

the refinement procedure. Since these differences turn out to be significant and systematic, the simulation results concerning the magnitude, anisotropy and anharmonicity of the motions are used to examine the source of the errors in the refinement.

Section 2 outlines the methods employed in this study. The approach used to generate the X-ray intensities from the simulation results and the details of the procedures employed for refining the data are described. The results are presented and analysed in section 3. Emphasis is placed on the atomic positions, the stereochemistry of the structure, and the atomic motions. The conclusions are outlined in section 4.

2. Methods

(a) *The calculation of diffraction intensities for a static structure*

For a perfect crystal with no thermal motion, the intensity of scattered X-rays is proportional to the square of the Fourier transform of the electron density in a unit cell (Woolfson, 1970). The Fourier transform of the electron density is called the structure factor, $F(\mathbf{Q})$, where $\mathbf{Q}$ is the scattering vector, defined by:

$$\mathbf{Q} = \frac{2\pi(\mathbf{e}-\mathbf{e}_0)}{\lambda}, \tag{1a}$$

where $\mathbf{e}_0$ and $\mathbf{e}$ are unit vectors along the directions of the wave vectors of the incident and scattered radiation, respectively, and λ is the wavelength of the X-radiation. In terms of the reciprocal lattice vectors, $\mathbf{a}^*$, $\mathbf{b}^*$ and $\mathbf{c}^*$, $\mathbf{Q}$ is given by:

$$\mathbf{Q} = 2\pi(h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*), \tag{1b}$$

where h, k and l are not, in general, required to be integral (Willis & Pryor, 1975). The structure factor, $F(\mathbf{Q})$, is thus given by:

$$F(\mathbf{Q}) = \int \mathrm{d}\mathbf{r}\, \rho(\mathbf{r})\, \mathrm{e}^{i\mathbf{Q}\cdot\mathbf{r}}, \tag{2}$$

where $\rho(\mathbf{r})$ is the electron density at $\mathbf{r}$ and the integral is over the unit cell. In most crystallographic applications, the molecular electron density is approximated by a superposition of the electron densities of the individual atoms. The electron density at a point $\mathbf{r}$ in the unit cell is then given by:

$$\rho(\mathbf{r}) = \sum_{i=1}^{N} \rho_i(\mathbf{r}-\mathbf{r}_i), \tag{3}$$

where the sum runs over the N atoms of the 1 or more molecules in the asymmetric unit of the unit cell, and $\rho_i(\mathbf{r}-\mathbf{r}_i)$ is the electron density at $\mathbf{r}$ due to an atom at $\mathbf{r}_i$. Substituting eqn (3) in eqn (2), we obtain:

$$\begin{aligned} F(\mathbf{Q}) &= \sum_{i=1}^{N} \int \mathrm{d}\mathbf{r}\, \rho_i(\mathbf{r}-\mathbf{r}_i)\, \mathrm{e}^{i\mathbf{Q}\cdot\mathbf{r}} \\ &= \sum_{i=1}^{N} f_i(\mathbf{Q})\, \mathrm{e}^{i\mathbf{Q}\cdot\mathbf{r}_i}, \end{aligned} \tag{4a}$$

where $f_i(\mathbf{Q})$ is the atomic scattering factor of the ith atom:

$$f_i(\mathbf{Q}) = \int \mathrm{d}\mathbf{r}\, \rho_i(\mathbf{r})\, \mathrm{e}^{i\mathbf{Q}\cdot\mathbf{r}}. \tag{4b}$$

The atomic electron densities are obtained from *ab-initio* quantum mechanical calculations and are Fourier transformed to obtain atomic scattering factors. These atomic scattering factors are, in general, complicated anisotropic expressions, and a further simplification is made by fitting a simple analytic, isotropic function to the *ab-initio* scattering factors. The most commonly used form is a sum of 2 to 4 Gaussians plus a constant. Defining $s = |\mathbf{Q}|/4\pi$, we have:

$$f(s) = \sum_{\alpha=1}^{n} a_\alpha\, \mathrm{e}^{-b_\alpha s^2} + c. \tag{5}$$

The parameters a_α, b_α and c are obtained by a least-squares fit to the *ab-initio* scattering factors. Parameters for a large number of atoms and ions are available for fits using 2 Gaussians (Moore, 1963) and 4 Gaussians (International Tables for X-ray Crystallography, 1974). We note that the electron density corresponding to the scattering factor in eqn (5) is a sum of Gaussians plus a delta function centred at the position of the atom.

At the current resolution limits of crystallographic data for proteins, eqns (4) and (5) are of sufficient accuracy (Moore, 1963) and so the structure factors can be computed readily as a sum of isotropic atomic scattering factors and phase factors (eqn (4)). Such a calculation, referred to as a direct summation, is very expensive for large molecules (Ten Eyck, 1973; Agarwal, 1978); it requires about 0·5 h on a VAX 11/780 computer for a 1·5 Å resolution calculation on myoglobin. We take an alternative approach, which is to use fast Fourier transform algorithms to calculate structure factors. Programs to do this have been available since about 1973 (Ten Eyck, 1973, 1977; Agarwal, 1978). Such a calculation proceeds in 2 steps. The first is the calculation of the electron density in the asymmetric unit of the molecule from a superposition of atomic electron densities and the construction of an electron density grid. In the second step, the structure factors are calculated by a finite discrete Fourier transform of the electron density grid using FFT† algorithms. In the available programs (Ten Eyck, 1977), the atomic electron densities used are obtained from the atomic scattering factors given by Moore (1963), which are in the form of 2 Gaussians plus a constant. Direct Fourier transformation of eqn (5) to obtain an expression for the electron density would introduce a delta function into the expression. To avoid this, a pseudo-temperature factor, B_0, is added to each atom. This pseudo-temperature factor can be scaled out of the structure factors after the Fourier transformation, and its inclusion is also important for reducing aliasing errors due to the discrete sampling of the electron density (Ten Eyck, 1977). The resulting model for the atomic electron density is:

$$\rho(r) = \sum_{\alpha=1}^{2} \frac{a_\alpha}{\sigma_\alpha^3} \exp\left[-\frac{\pi r^2}{\sigma_\alpha^2}\right] + \frac{c}{\sigma_3^3} \exp\left[-\frac{\pi r^2}{\sigma_3^2}\right], \tag{6}$$

where a_α, b_α and c are the same as in eqn (5) and $\sigma_\alpha^2 = (b_\alpha + B_0)/4\pi$ and $\sigma_3^2 = B_0/4\pi$. These methods are described in detail by Ten Eyck (1973, 1977) and by Agarwal (1978), and are not discussed further here.

(b) *X-ray intensities from a molecular dynamics simulation*

X-rays scattered from a crystal can be considered as having 2 components. One of them, usually referred to as the Bragg scatter, exists only at scattering angles that satisfy Bragg's law and gives rise to the discrete spots observed in diffraction photographs (Willis & Pryor,

† Abbreviations used: FFT, fast Fourier transfer; r.m.s., root-mean-square; ps, picosecond.

1975; and see the Appendix). The intensity of the Bragg scatter is proportional to the square of the Fourier transform of the average electron density in a unit cell (Stewart & Feil, 1980). The other component is not restricted to the reciprocal lattice points and is referred to as thermal diffuse scatter (Willis & Pryor, 1975; Amoros & Amoros, 1968), and to compute this would require information about the correlations of the motions of atoms in one unit cell with those in another. Diffuse scatter arises also due to disorder in the crystal (Amoros & Amoros, 1968) and, though observed in protein crystals (Wilson *et al.*, 1983), its effects are generally ignored in processing and analysing the data. The only way in which the effects of thermal motion on X-ray diffraction data are included for protein crystals is by assuming that the average electron density associated with a given atom is not that obtained if the atom were fixed in position (eqn (6)); instead, it is given by a convolution of the fixed density with a positional probability distribution function arising from the motion of the atom. These distribution functions, also called thermal smearing functions (Willis & Pryor, 1975), are the Fourier transforms of the atomic Debye–Waller factors.

In the computer "experiment" to be described here, we calculate average intensities by using a molecular dynamics simulation, which yields a trajectory that gives the position of every atom in the protein as a function of time (Karplus & McCammon, 1981, 1983). The electron density for a given co-ordinate set is calculated by use of eqn (6) for each atom and the resulting electron densities are averaged over the co-ordinate sets from the trajectory. It is important to note that no assumed model for the probability distribution functions of the atomic motions is used in this calculation. The averaging of the electron densities over the trajectory corresponds to convoluting the static electron density with the probability distribution functions obtained from the simulation. This is equivalent to calculating structure factors from co-ordinate sets sampled from the simulation and averaging them:

$$I(\mathbf{Q}) \propto |\int \mathrm{d}\mathbf{r}\langle\rho(\mathbf{r})\rangle\, \mathrm{e}^{i\mathbf{Q}\cdot\mathbf{r}}|^2 = |\langle F(\mathbf{Q})\rangle|^2. \quad (7)$$

Equation (7) is valid if it is assumed that there is no correlation in the motions of different protein molecules in the crystal (see the Appendix). This assumption, which corresponds to neglecting thermal diffuse scattering, is the standard one made in crystal structure refinements (Willis & Pryor, 1975). An alternative limiting assumption would be that all protein molecules are moving in phase, i.e. that at every instant the molecules in the crystal are all identical and that they all evolve identically in time (Appendix). The correlated motion assumption yields an intensity of the form:

$$I(\mathbf{Q}) \propto \langle |F(\mathbf{Q})|^2 \rangle. \quad (8)$$

We calculated structure factors between 10·0 Å and 1·5 Å using eqns (7) and (8). The crystallographic R-factor, R, is commonly used to indicate the quality of agreement between 2 sets of structure factors:

$$R = \frac{\sum_{\mathbf{Q}} ||F_o(\mathbf{Q})| - |F_c(\mathbf{Q})||}{\sum_{\mathbf{Q}} |F_o(\mathbf{Q})|}. \quad (9)$$

The R-factor between structure factors calculated from a 25 ps simulation of myoglobin using eqns (7) and (8) is 36% (Appendix). This is obviously an extreme case, but does suggest that correlations may sometimes be important. Unfortunately, the intermediate case is difficult to treat for proteins, though it may be approached by analysis of the lattice modes of crystals. As stated earlier, we follow standard practice and neglect all correlations between unit cells in calculating structure factors; i.e. we use eqn (7) for the calculation of data described in this paper to approach most closely the procedure usually followed in protein X-ray refinements.

We decided to do our computer experiment at a resolution of 1·5 Å as this is comparable with the resolution of the best X-ray data currently available for proteins the size of myoglobin (Kuriyan *et al.*, unpublished results; Phillips, 1980). The myoglobin molecule was placed in a crystal lattice of monoclinic system with the symmetry of space group $P2_1$, and with 1 molecule in the asymmetric unit. The unit cell was assumed to have parameters $a = 64{\cdot}31$ Å, $b = 30{\cdot}85$ Å, $c = 34{\cdot}85$ Å, $\alpha = 90{\cdot}0°$, $\beta = 105{\cdot}85°$ and $\gamma = 90{\cdot}0°$, corresponding to the experimental parameters for met-myoglobin at 300 K (Hartmann *et al.*, 1982). The structure factors also depend on the orientation and position of the molecule in the unit cell. To make the situation comparable to the experimental one, the average structure from the dynamics was superimposed, by least-squares, on the experimental structure at 300 K. A small translation of 0·352 Å along **a**, −0·006 Å along **b** and 0·117 Å along **c** was obtained and applied to all the co-ordinate sets sampled from the simulation.

Given this space group, unit cell and orientation of the molecule in the unit cell, there are about 22,000 unique structure factors (not including Friedel pairs) between 10·0 Å and 1·5 Å. Calculation of these structure factors for one structure from the simulation using FFTs takes about 5 min, which is about 6 times faster than the direct summation. However, much greater savings in computer time can be achieved in calculating the averaged structure factors because the electron density calculation is fast (about 1·5 min per structure) and, instead of averaging structure factors, we can average the electron density and then do just 1 Fourier transform at the end.

A program written by L. F. Ten Eyck was used (Ten Eyck, 1977). It employs eqn (6) to calculate atomic electron densities; they are superimposed to obtain the molecular electron density, from which the electron density in the asymmetric unit of the unit cell is calculated. The electron density was sampled on a grid with 160 grid points along **a**, 88 along **b**, and 88 along **c**. In the $P2_1$ space group, only half of the unit cell along **b** needs to be included. The sampling intervals used are finer than the recommended interval of 1/3 the minimal interplanar spacing of the data (0·5 Å in this case). This, along with a pseudo-temperature factor of 20 Å^2 added to each atom (see eqn (6)), serves to reduce errors due to finite sampling (Ten Eyck, 1977). Structure factors were calculated from the electron density grid by use of a set of FFT subroutines written by L. F. Ten Eyck and modified by G. Bricogne (Ten Eyck, 1973, 1977).

The refinement program used in this work, PROLSQ (Konnert & Hendrickson, 1980), uses a 4-Gaussian form for the atomic scattering factors rather than the 2-Gaussian form built into the FFT calculations. The 2-Gaussian form is preferable in the electron density calculation as it leads to significant enhancement in the speed of the calculation. The refinement program also computes structure factors by direct summation (eqn (4)) as the derivatives are more easily obtainable this way. To estimate errors introduced by these discrepancies between the 2 programs, R-factors between structure factors

calculated using 2 and 4 Gaussians and by the FFT program and by direct summation were calculated. The *R*-factor between the FFT and direct summation structure factors using a 2-Gaussian electron density model in both cases is 0·64%. With a 4-Gaussian model in the direct summation and a 2-Gaussian model in the FFT, the *R*-factor is 0·79%. These errors are negligible. Thus it was concluded that no significant advantage would be gained by using the 4-Gaussian form in the electron density calculation.

(c) *The simulation used and the X-ray data sets generated*

Most of the work reported here is based on a 25 ps segment of a 50 ps simulation of myoglobin. Some results are presented for refinement of data generated from a 300 ps simulation of myoglobin (Levy *et al.*, 1985). Both simulations were calculated with a version of the CHARMM program (Brooks *et al.*, 1983) and used identical initial structures and parameters. The initial co-ordinates for the simulation were obtained from a refined crystal structure of met-myoglobin at 250 K (Frauenfelder *et al.*, 1979). The model for the protein included hydrogen atoms only for methyl groups and the total system simulated included 1423 atoms (1217 non-hydrogen protein atoms, 162 methyl hydrogens, 43 heme atoms and 1 water molecule bound to the iron). No solvent molecules were included. The average temperature of the simulations was 298 K (Levy *et al.*, 1985). No hydrogen atoms were included in the structure factor calculations.

The 25 ps simulation was sampled at intervals of both 0·25 ps and 0·05 ps. While the simulation was sampled every 0·25 ps, the structure factors were calculated and averaged in 2 different ways, as an internal check. In one case, structure factors were calculated for all the 100 co-ordinate sets from the trajectory, i.e. 100 electron density calculations and Fourier transformations were done to obtain the modulus of the complex mean structure factors. In the other case, the structure factors were calculated by averaging the electron density and doing only 1 Fourier transform on the averaged electron density. The *R*-factor between structure factors calculated in the 2 different ways is 1·6% (at 1·5 Å resolution). Each FFT introduces a small error due to finite sampling of the electron density (Ten Eyck, 1977) and the 2nd method, which involves only 1 Fourier transform instead of 100, is expected to be more accurate. However, the differences found here are small compared to the final *R*-factors of the refined structures. The structure factors calculated using 100 FFTs were used in 2 of the refinements that follow and will be referred to as the $|\langle F\rangle|^{0.25}$ set.

The 25 ps simulation was also sampled every 0·05 ps, and the averaged electron density obtained from the 500 co-ordinate sets was used to generate structure factors. This set is referred to as the $|\langle F\rangle|^{0.05}$ data set. Comparison of the $|\langle F\rangle|^{0.25}$ and $|\langle F\rangle|^{0.05}$ sets allows us to check that the results obtained using the $|\langle F\rangle|^{0.25}$ set were not biased by poor sampling of the trajectory. The $|\langle F\rangle|^{0.05}$ data set was used for 4 of the refinements reported below. The *R*-factor between data calculated using sampling intervals of 0·25 ps and 0·05 ps is 3·2% (at 1·5 Å resolution). Thus a small error is introduced by the coarser sampling, but again this is well below the final *R*-factors of any of the refinements. Finally, the entire 300 ps of the longer trajectory was sampled every 0·25 ps, and the average electron density was determined from 1200 co-ordinate sets sampled from the simulation. Structure factors were calculated by Fourier transformation of the average density and this data set, which is referred to as the $|\langle F\rangle|^{300}$ set, was used for 1 of the refinements reported below.

(d) *Modelling thermal motion in crystallographic refinement*

As mentioned earlier, most refinements of protein structures made to date assume a harmonic, isotropic model for the probability distribution functions. This leads to the following expression for the time-averaged structure factors (Willis & Pryor, 1975):

$$\langle F(\mathbf{Q})\rangle = \sum_{j=1}^{N} f_j(\mathbf{Q})\, e^{i\mathbf{Q}\cdot\langle \mathbf{r}_j\rangle}\, e^{W_j(\mathbf{Q})}, \qquad (10)$$

where, as before, $f_j(\mathbf{Q})$ is the atomic scattering factor and $\langle \mathbf{r}_j\rangle$ is the average position of the jth atom. The term $e^{W_j(\mathbf{Q})}$ is the atomic Debye–Waller factor and, in the isotropic case, $W_j(\mathbf{Q})$ is given by:

$$W_j(\mathbf{Q}) = -\tfrac{1}{6}\langle \Delta\mathbf{r}_j^2\rangle |\mathbf{Q}|^2 = -\tfrac{8}{3}\pi^2\langle \Delta\mathbf{r}_j^2\rangle s^2, \qquad (11)$$

where $\langle \Delta\mathbf{r}_j^2\rangle$ is the mean-square fluctuation of the jth atom and $s = |\mathbf{Q}|/4\pi$. The term $(8/3)\pi^2\langle \Delta\mathbf{r}_j^2\rangle$ is referred to as B_j, the atomic *B*-factor or temperature factor (Willis & Pryor, 1975). Equation (10) is a type of model very different from the approach used in the dynamics, in that instead of a full averaging of the atomic electron density, as in the dynamics, averages of the position, $\langle \mathbf{r}_j\rangle$, and the mean-square fluctuation, $\langle \Delta\mathbf{r}_j^2\rangle$ σ_j^2, are introduced. The average intensity at a reciprocal lattice point, which is what is measured, is proportional to the absolute value, squared, of the structure factor (eqn (7)):

$$I(\mathbf{Q}) \propto |\langle F(\mathbf{Q})\rangle|^2. \qquad (12)$$

(e) *Least-squares refinement*

The standard crystallographic refinement process iteratively improves the agreement between the structure factors calculated from a model structure and those derived from the measured X-ray intensities. This is done by varying parameters in the model based on solutions to the linearized least-squares formulation of the problem (Konnert, 1976; Hendrickson, 1985). The function Φ minimized is of the form:

$$\Phi = \sum_{\mathbf{Q}} w(\mathbf{Q})\,||F_o(\mathbf{Q})| - |F_c(\mathbf{Q})||^2, \qquad (13)$$

where $|F_o|$ is the experimental amplitude of the structure factor and $|F_c|$ is that calculated from the model (eqn (10)). $w(\mathbf{Q})$ is the weight assigned to the structure factor. As discussed below, eqn (13) is generally modified in protein refinements to allow for the introduction of restraints on the structure.

All the refinements reported here were done using the restrained-parameter, least-squares program PROLSQ (Konnert & Hendrickson, 1980). Four parameters were refined for every non-hydrogen atom in met-myoglobin: 3 Cartesian co-ordinates and 1 isotropic temperature factor. The neglect of hydrogen atoms is not an approximation in this work, as none was included in the calculation of structure factors.

The initial model used in all cases was the averaged structure from the molecular dynamics simulation. The average co-ordinates obtained by sampling at 0·05 ps and 0·25 ps were identical to within 0·01 Å, and so just 1 structure was used as the initial model for all the

refinements of the 25 ps data. The dynamical average structure obtained by sampling the simulation every 0·25 ps was used as the initial structure for refinement of the 300 ps data. A uniform temperature factor was assigned to all the atoms at the start of the refinements.

(f) *The use of restraints in refinement*

While refining a protein structure, it is usually the case that the number of structure factors experimentally measured is insufficient to ensure that the refinement will be well-behaved at the desired resolution. This point has been discussed in detail by Konnert & Hendrickson (1980), who include stereochemical data as additional information available to the refinement program. Table 1 shows the ratio of parameters to observables for myoglobin at various minimal interplanar spacings for various numbers of refinement parameters. From this Table, it can be seen that for isotropic *B*-refinement at 1·5 Å resolution, the number of independent data points exceeds the number of variable parameters by a factor of 4·4. Thus it should be possible to refine co-ordinates and isotropic temperature factors for each atom without necessarily resorting to restraints. Though this is shown to be possible in this work, it is much more difficult with experimental data, where errors in the measurements often limit the number of reliable data.

To incorporate stereochemical restraints, PROLSQ requires a dictionary of ideal amino acid structures. The stereochemical restraints include 1–2 distance restraints for the bonds, 1–3 distance restraints for the angles, planarity restraints, torsional restraints and non-bonded contact restraints (Hendrickson, 1980; Hendrickson & Konnert, 1980; Hendrickson, 1985). In addition to these, the variation of the temperature factors of atoms that are bonded to each other or to the same 3rd atom is restrained to lie within specified values (Konnert & Hendrickson, 1980). Finally, the program also restrains the calculated shifts in the parameters. The restraints are added to the observational equation as additional "observations", and so the function that is minimized is:

$$\Phi = \sum_{\mathbf{Q}} w(\mathbf{Q})\|F_o(\mathbf{Q})|-|F_c(\mathbf{Q})\|^2 + \sum_d w_d\Delta^2, \qquad (14)$$

where $w(\mathbf{Q})$, the weight assigned to the structure factors, varies linearly with $|\mathbf{Q}|$, so that low-resolution structure factors are weighted more than high-resolution structure factors (eqn (14): Hendrickson, 1980):

$$w(\mathbf{Q}) = \alpha - \beta(s-\gamma), \qquad (15)$$

with $s = |\mathbf{Q}|/4\pi$. The value of α controls the weight assigned to the structure factor data relative to the restraint terms as a whole, while β controls the weight assigned to the higher-resolution data relative to the lower-resolution data. γ is the value of s at which the line defined by eqn (15) is pivoted, and a value of 1/6 $Å^{-1}$ was used in all the refinements described here (Hendrickson & Konnert, 1980; Hendrickson, 1980). The values of α and β are set by trial and error during the course of a refinement (Hendrickson, 1980). Δ is the deviation of a restrained parameter from its ideal value, and w_d is the weight assigned to the restraint. The weights on restraints used in this work are listed in Table 2A, and these are similar to the values suggested by Hendrickson (1980), and those used in previous refinements of myoglobin (Frauenfelder *et al.*, 1979; Hartmann *et al.*, 1982; Kuriyan *et al.*, unpublished results).

The weights can be thought of as the inverse of the expected variance of Δ, and their values can be changed to "tighten" or "loosen" any particular stereochemical restraint. In practice, these variances are used as target values for the observed values of Δ, and the overall weights on the restraints are varied so as to make the refined structure conform to these target values

Table 1

The number of independent reflections for myoglobin as a function of resolution

Resolution (Å)	Number of independent data points	Diffraction data per variable parameter: *X*, *Y*, *Z* thermal ellipsoid	*X*, *Y*, *Z* *B*	Dihedrals only
3·0	~3000	—	0·6	5
2·0	~10,000	—	2·0	15
1·5	~22,000	1·9	4·4	
1·2	~40,000	3·5	7·9	
0·86	~125,000	11·0	24·8	

Table 2

Restraint parameters and results

A. Weights on the various classes of restraints

Type of restraint	Target standard deviation
1–2 Bond distances (Å)	0·030
1–3 Angle distances (Å)	0·040
1–4 Planar distances (Å)	0·052
Planar groups (Å)	0·025
Chiral groups ($Å^3$)	0·150
Temperature factor restraints ($Å^2$):	
Backbone bonded pairs	1·0
Backbone angle pairs	1·5
Side-chain bonded pairs	1·0
Side-chain angle pairs	1·5
Non-bonded contact restraints:	
Non-bonded contact pairs (Å)	0·5
Torsional restraints:	
Torsion angles (deg.)	10·0
Positional shift restraints (Å)	0·3
B-factor shift restraint ($Å^2$)	3·0
Occupancy factor shift restraint	0·05

B. *Deviations from ideal PROLSQ stereochemistry (25 ps simulation)*

Type	Statistics from the average dynamics structure: Average deviation†	r.m.s.	Statistics done along the simulation: Average deviation†	r.m.s.
Distances (Å)				
Backbone bonds	−0·02	0·049	0·0007	0·015
Carbonyl O bonds	−0·06	0·100	−0·0083	0·009
Side-chain bonds	−0·11	0·200	0·002	0·023
Backbone angles (1–3 distances)	−0·019	0·055	0·014	0·040
Carbonyl O angles (1–3 distances)	−0·090	0·150	−0·020	0·020
Side-chain angles (1–3 distances)	−0·140	0·270	−0·005	0·040
Angles (deg.)				
Backbone angles	−2·39	4·05	−1·234	2·12
Side-chain angles	−3·90	12·72	1·090	3·31

† Over the molecule.

(Hendrickson, 1980). This can be done by varying the relative weights of the structure factors and the restraints during the course of the refinement by changing the value of α in eqn (14), thus controlling how much the model is forced to adhere to the restraints. One indication of the tightness of the restraints is the deviation of 1–2, 1–3 and 1–4 distances from their ideal values. A 1–2, 1–3 or 1–4 distance for which $|\Delta|$ is more than $2w_d^{-1/2}$ shall be referred to as a deviant distance. The total number of restrained distances is 3500.

The ideal values for the stereochemical parameters are derived from crystal structures of small molecules. The large amplitude motions observed in protein crystals, both in simulation results and in refined temperature factors (Karplus & McCammon, 1981, 1983; Petsko & Ringe, 1984) might be expected to cause the average structure to exhibit deviant stereochemistry (Karplus, 1981). The bias introduced by restraining parameters is examined by doing refinements with loose restraints, tight restraints and no restraints on the co-ordinates and temperature factors. Another possible cause for deviations between the stereochemical parameters in the dynamics average structure and the ideal structure is that the parameters used in the molecular dynamics simulation, which determine the equilibrium values of the individual internal co-ordinates of the molecule, are different from those used in the refinement. This point is discussed further in Results and Discussion.

(g) *The refinements*

(i) *Refinement against* $|\langle F\rangle|^{0.25}$ *data with loose restraints*

The initial atomic positions used as input to the refinement program were the average positions from the simulation. Only co-ordinates were refined for the initial model at 2·0 Å resolution with a constant overall temperature factor of 2·0 Å^2. The *R*-factor dropped from 37% to 23·7% in 7 cycles. At this stage, the overall temperature factor was set to 13·5 Å^2, which is the value at which the *R*-factor is at a minimum, and the refinement was continued at 2·0 Å resolution. Merely increasing the temperature factors from 2·0 Å^2 to 13·5 Å^2 lowered the *R*-factor from 23·7% to 19·2%. Then 8 further cycles of co-ordinate and individual temperature factor refinement lowered the *R*-factor to 13·0%, at which stage the refinement had reached apparent convergence.

Initially, tight restraints were kept on the stereochemistry and the number of deviant distances dropped to 171 from 1093 (see Table 3). For the 1·5 Å resolution refinement, where the ratio of observations to variable parameters was greater, the weights on the stereochemical restraints were relaxed and kept very loose (by increasing the value of α in eqn (15), and the number of deviant distances increased at every cycle. In 8 cycles, the *R*-factor had dropped to 13·6% with 1372 deviant distances. At this point, the *R*-factor did not change on continuing the refinement, and the co-ordinates and temperature factors were saved for analysis. The *R*-factors in various resolution shells for the different refinements are reported in Table 4. Final co-ordinates and temperature factors from this set will be referred to as the $\langle F\rangle_{\text{restr}}^{0.25}$ set.

(ii) *Refinement against* $|\langle F\rangle|^{0.25}$ *data with no restraints*

In this case, the weights on all stereochemical and temperature-factor restraints were set to zero. The refinement was started by including all the data between 10·0 Å and 1·5 Å, and the initial model was the dynamics structure with an overall temperature factor of 15·0 Å^2. The initial *R*-factor was 37·0% at 1·5 Å. The number of deviant distances increased to about 2000 in a few steps and then stayed more or less constant throughout the refinement. In 14 cycles, the refinement reached apparent convergence at an *R*-factor of 13·6%, with 2035 deviant distances (see Table 3). Co-ordinates and temperature factors from this refinement will be referred to as the $\langle F\rangle_{\text{unrestr}}^{0.25}$ set.

(iii) *Refinement against* $|\langle F\rangle|^{0.05}$ *data with loose restraints*

The refinement was started with an overall temperature factor of 15 Å^2 and initially only data to 2·0 Å were included. Ten cycles of least-squares refinement reduced the *R*-factor to 10·5% with 1326 deviant distances. At this stage, data to 1·5 Å were included and 9 more cycles of refinement dropped the *R*-factor to 13·6%, with 1391 deviant distances. Two

Table 3

Overall statistics for the seven refinements

Description of structure	*R*-factor (%)	Deviant distances	r.m.s. Δ of			
			Bonds	Angles	1–4 dist.	Planes
M.D. (25): average structure and fluctuations from 25 ps dynamics	19·2	1093	0·146	0·204	0·160	0·035
Restr1: $\langle F\rangle^{0.25}$ with loose restraints	13·6	1372	0·089	0·105	0·099	0·051
Restr2: $\langle F\rangle^{0.05}$ with loose restraints	13·6	1391	0·080	0·105	0·105	0·050
Trestr2: $\langle F\rangle^{0.05}$ with tight restraints	16·6	157	0·026	0·042	0·043	0·018
Unrestr1: $\langle F\rangle^{0.25}$ with no restraints	13·6	2035	0·175	0·210	0·186	0·095
Unrestr2: $\langle F\rangle^{0.05}$ with no restraints	12·9	1865	0·171	0·211	0·172	0·112
Altconf: $\langle F\rangle^{0.05}$ alternate conformations 300 ps	12·6	1478	0·090	0·111	0·109	0·060
with loose restraints	21·5	1759	0·098	0·134	0·141	0·055
Target stereochemical standard deviations			0·030	0·040	0·052	0·025

Table 4

The final refined R*-factors between various resolution limits are given for refinements of the* $\langle F \rangle^{0.05}$ *data set from the 25 ps simulation*

	R-factors (%) in shells of resolution Resolution (Å)							
	5·00	4·00	3·20	2·50	2·00	1·75	1·50	Overall
Restrained refinement	9·6	8·0	8·0	9·9	13·2	18·0	20·4	13·6
Unrestrained refinement	7·3	6·0	6·6	9·4	12·9	17·7	21·7	12·9
Restrained refinement with alternate conformations	7·0	6·2	6·4	8·5	11·9	17·4	21·0	12·6

further cycles of refinement resulted in no change in the *R*-factor and the refinement was stopped. Final co-ordinates and temperature factors from this refinement will be referred to as the $\langle F \rangle^{0.05}_{\text{trestr}}$ set.

(iv) *Refinement against* $|\langle F \rangle|^{0.05}$ *data with tight restraints*

This refinement was a continuation of the previous one with greater weights on the restraints. The *R*-factor increased from 13·6% to 16·6% but the number of deviant distances dropped from 1391 to 157. This structure will be referred to as the $\langle F \rangle^{0.05}_{\text{trestr}}$ structure.

(v) *Refinement against* $|\langle F \rangle|^{0.05}$ *data with no restraints*

Refinement was started at 1·5 Å and 14 cycles dropped the *R*-factor to 12·9% from 37%. Three further cycles of refinement resulted in no change in the *R*-factor, and the final structure, with 1865 deviant distances, will be referred to as the $\langle F \rangle^{0.05}_{\text{unrestr}}$ structure.

(vi) *Refinement against* $|\langle F \rangle|^{0.05}$ *data with alternate conformations*

Difference electron density maps with coefficients $(2F_o - F_c)\exp(i\alpha_c)$ (Blundell & Johnson, 1976) were examined on an Evans and Sutherland PS300 graphics system using the software FRODO (Jones, 1982). The phases, α_c, and the model structure factors, F_c, were calculated from the $\langle F \rangle^{0.05}_{\text{restr}}$ structure. F_o is the amplitude of the observed structure factor, i.e. the "experimental" data obtained from the simulation. On the basis of the difference maps alone, 10 residues showed clear indications of conformational disorder in their side-chains, and these were modelled by 2 conformations for each of the residues. In all cases, only 2 conformations were built, and these differed only from the C_γ atom outwards.

Only one variable occupancy factor was refined for each residue with alternate conformations. All the atoms belonging to one conformation were constrained to have the same occupancy and the occupancies of the 2 alternate conformations were constrained to add up to 1·0. The refinement was started with equal weights and temperature factors assigned to all atoms with alternate conformations. All other atoms had the same parameters as at the end of the $\langle F \rangle^{0.05}_{\text{restr}}$ refinement. Stereochemical restraints were applied to all the atoms and refinement was started at 2·0 Å to allow the atomic positions to adjust. Nine cycles of refinement lowered the *R*-factor from 11·5% to 8·0%, with 1452 deviant distances. Data to 1·5 Å were included at this point, and a further 8 cycles of refinement lowered the *R*-factor from 14·1% to 12·6%. The *R*-factor would not drop on continuing the refinement, and the process was stopped. The co-ordinates and the temperature factors from this refinement will be referred to as the $\langle F \rangle^{0.05}_{\text{altconf}}$ set.

(vii) *Refinement against 300 ps data with loose restraints*

The initial model structure was the average structure from the 300 ps trajectory with a uniform temperature factor of 15 $Å^2$ assigned to each atom. The first few cycles of refinement included data between 10·0 Å and 2·0 Å, this was later extended to include all the data between 10·0 Å and 1·5 Å. Twenty-one cycles of refinement with very loose restraints finally reduced the *R*-factor to 21·5%, at which point the *R*-factor would not drop further. The refinement was stopped and co-ordinates and temperature factors saved for analysis; they are referred to as the $\langle F \rangle^{300}$ set.

3. Results and Discussion

(a) *The* R*-factors of the refined structures*

The *R*-factor (eqn (9)) is the most commonly used indicator of the quality of a refined model and the final *R*-factors of the six refinements of the 25 ps data are compared in Tables 3 and 4. The *R*-factors range from 12·6% for the refinement with loose restraints and ten alternate conformations, to 16·6% for the refinement with tight restraints. Refinements with both loose restraints and with no restraints lead to similar *R*-factors, while tight restraints increase the *R*-factor by about 3%. On comparing these values with the *R*-factor of 19·2% obtained using the average dynamics structure (with isotropic atomic temperature factors calculated from the exact mean-square fluctuations from the simulation, using eqns (10) and (11)), we see that the average dynamics structure and exact, isotropic fluctuations do not yield the best fit to the structure factor data.

The refined *R*-factors are all higher than the experimental refined *R*-factors for small molecules, which are usually less than 5% (Dunitz, 1979), but they are comparable with experimental refined *R*-factors found for most proteins at 1·5 Å resolution. A recent X-ray refinement of CO-myoglobin, for example, resulted in *R*-factors of 16·5% for a structure with loose restraints and several alternate conformations and 17·1% for a structure with tight restraints (Kuriyan *et al.*, unpublished results). Comparisons of *R*-factors should be treated with

Table 5

B-factors for various crystal structures of myoglobin compared with molecular dynamics values

Source	Backbone average temperature factor	Side-chain average temperature factor
Met-myoglobin (Frauenfelder *et al.*, 1979)	11·8	13·1
Oxy-myoglobin (Phillips, 1980)	11·5	21·1
25 ps molecular dynamics	12·3	26·8
25 ps molecular dynamics restrained refinement	11·3	16·5
25 ps molecular dynamics refinement with tight restraints	12·5	14·5
25 ps molecular dynamics unrestrained refinement	11·7	17·6
300 ps molecular dynamics	25·6	48·6
300 ps molecular dynamics restrained refinement	16·8	21·1
300 ps molecular dynamics restrained refinement with higher initial *B*	19·2	23·5

caution, however, because the *R*-factor depends on the resolution of the data (usually increasing with higher-resolution data) and the ratio of observables to parameters. Decreasing the ratio of observables to variable parameters, either by increasing the complexity of the refinement model or by not including all the unique structure factor data at a particular resolution (usually because of experimental uncertainty), generally results in a decrease in the *R*-factor (Hirshfeld & Rabinovich, 1973). The CO-myoglobin model (Kuriyan *et al.*, unpublished results) was refined against only 10,449 unique structure factors between 10·0 Å and 1·5 Å with an observations to parameters ratio of 2·07, whereas in the refinements reported here all the 21,942 unique structure factors between the same resolution limits were included resulting in an observations to parameters ratio of 4·35. The *R*-factors from the two refinements are therefore not strictly comparable; the *R*-factors reported in this work are expected to be higher than those that would be obtained from refinements done on a smaller subset of the same structure factors (Hirshfeld & Rabinovich, 1973).

The high *R*-factors that are the converged limits of all the refinements appear to be due to the neglect of the anharmonic, anisotropic nature of the atomic fluctuations in the refinement program. That they are not due to any kind of problem with the least-square algorithm itself is demonstrated both by the high *R*-factor of the average dynamics structure with exact *B*-factors from the simulation and also by test refinements carried out on structure factor data generated from structures with isotropic *B*-factors (Kuriyan, unpublished results). Refinement against test data generated from single structures with isotropic and harmonic fluctuations always converge rapidly to a low *R*-factor (less than 1·0%) and yield refined structures virtually identical with the structures used to generate the data. The refined temperature factors are sometimes in error by a constant amount, but this is detected easily and can be corrected by calculating the *R*-factor as a function of overall shifts in the *B*-factor (Kuriyan, unpublished results).

The final *R*-factor of 21·5% for the refinement of the 300 ps data is significantly higher than any of the *R*-factors for the 25 ps data. The mean-square fluctuations of the atoms over the 300 ps period are roughly twice as great as the fluctuations over a 25 ps period (see Table 5) and the structure seems to undergo some larger-scale changes over the longer period (Levy *et al.*, 1985); thus, the higher *R*-factor is probably due to the isotropic, harmonic model being even less applicable to the simulation results for the long time-scale. We have not investigated whether the *R*-factor can be lowered by modelling disordered regions of the protein.

(b) *Restraints on stereochemistry and temperature factors*

The bond and angle terms are the most important of the stereochemical restraints. In PROLSQ they are both given as distance restraints, 1–2 bonded distances being restrained for the bonds and 1–3 angle distances being restrained for the angles. To assess the effect of dynamics on the ideality of the bonds and angles, the distances were calculated from the simulation in two different ways. In the first case, the distances were calculated from the average dynamics structure, i.e.:

$$\langle d_{ij} \rangle = |\langle \mathbf{r}_i \rangle - \langle \mathbf{r}_j \rangle|, \tag{16}$$

and in the second case the distances were calculated from structures sampled every 0·05 ps from the simulation, and then averaged:

$$\langle d_{ij} \rangle = \langle |\mathbf{r}_i - \mathbf{r}_j| \rangle. \tag{17}$$

The average deviations and r.m.s. deviations of the distances for various classes of atoms from their ideal values (as defined by the PROLSQ dictionary) are shown in Table 2. It is seen that the average and the r.m.s. of the deviations from ideality for all distances are larger for the average structure than for structures sampled from the simulation. The average deviations indicate that the bonds and angles are systematically smaller in the average structure than in the ideal dictionary. The deviations are smallest, in both cases, for the backbone atoms, and largest for atoms more than two atoms along the side-chain. Comparing the r.m.s. deviations for side-chain distances with the weights used in refinement, r.m.s. deviations in the average structure are about a factor of 10 higher than the standard deviations implicit in the weights. The deviations from ideality for the two different averages were calculated explicitly for the bond angles (instead of just the 1–3 distance) and these results are also given in Table 2.

Figure 1(a) shows the deviations, for both kinds

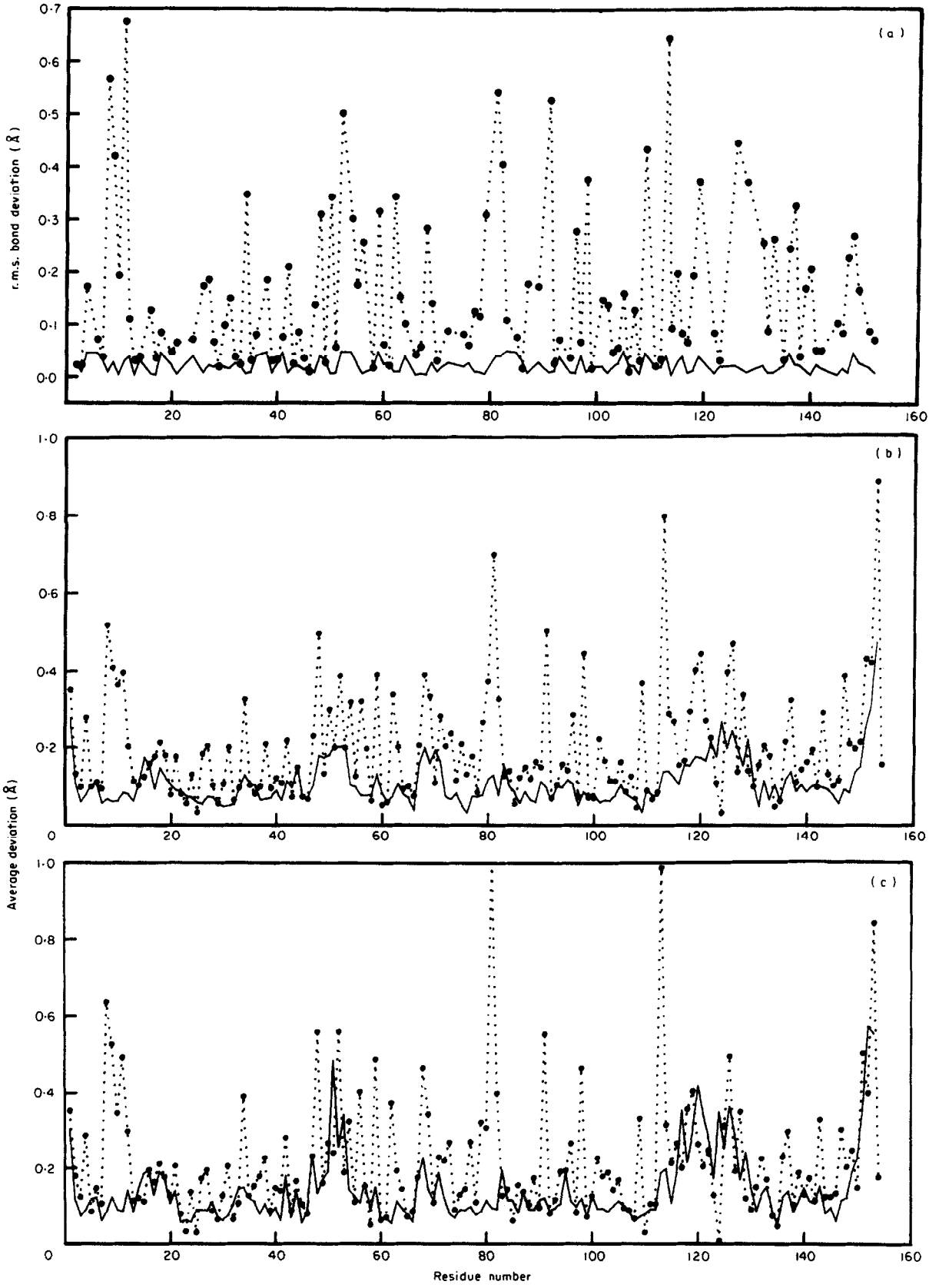

Figure 1. (a) Deviation of side-chain bonds from PROLSQ ideal values. The average bond length d_{ij} is calculated in 2 ways, sampling the 25 ps simulation every 0·05 ps: (1) $d_{ij} = |\langle \mathbf{r}_i \rangle - \langle \mathbf{r}_j \rangle|$ (dotted line); (2) $d_{ij} = \langle |\mathbf{r}_i - \mathbf{r}_j| \rangle$ (continuous line). (b) and (c) Positional error in refinement of the 25 ps data: the deviations between atomic positions are calculated after the molecular dynamics average structure and the refined structures are superimposed by least-squares. The deviations for backbone atoms (N, C_α and C; continuous line) and side-chain atoms (dotted line) are averaged over residues. Deviations are shown for (b) a restrained refinement ($\langle F \rangle^{0\cdot05}_{\text{restr}}$) and (c) for an unrestrained refinement ($\langle F \rangle^{0\cdot05}_{\text{unrestr}}$).

of averages, as a function of residue number for bonds between side-chain atoms. Except for one or two residues, both averages yield uniformly small deviations (less than about 0·04 Å) for backbone bonds. For side-chain bonds, the deviations calculated from averages over the simulation are still uniformly small, but those calculated from the average structure have a much larger deviation, ranging from about 0·1 Å to 0·8 Å (see Fig. 1(a)). The result for bond angles are very similar, with backbone angles deviating less than 3° to 4° for both averages, but with side-chain angles deviating as much as 40° to 50° in the average structure.

In all cases the r.m.s. deviations in stereochemistry sampled from the simulation are no larger than the target r.m.s. deviations from ideality (see Table 2A), despite the fact that individual structures exhibit sizeable fluctuations in the internal co-ordinates. This indicates that the equilibrium values for bonds, angles, etc. in the CHARMM potential (Brooks *et al.*, 1983) are not significantly different, as far as restraints in the refinement are concerned, from the ideal values in the PROLSQ dictionary. As expected, the regions of the protein with large deviations in geometry in the average structure correlate well with regions of the protein with high mobility in the simulation (cf. Fig. 5).

It is clear from this analysis that if there are large-scale motions occurring in the protein it may be inappropriate to impose strict stereochemical restraints (Karplus, 1981; Yu *et al.*, 1985). If the average dynamical structure is considered to be the correct structure, then refinement with large weights on the stereochemical restraints would clearly lead to deviations from the average structure. The r.m.s. values of Δ (see eqn (14)) for four of the most important classes of restraints are shown in Table 3 for the average dynamical structure as well as for the various refined structures. For 1–2, 1–3 and 1–4 distance restraints and for planarity restraints, the r.m.s. values of Δ are 0·15 Å, 0·20 Å, 0·16 Å and 0·04 Å, respectively, in the average structure, as compared to 0·08 to 0·09 Å, 0·11 Å, 0·10 to 0·11 Å and 0·05 Å in the loosely restrained structures, and 0·17 to 0·18 Å, 0·21 Å, 0·17 to 0·19 Å and 0·10 to 0·11 Å in the unrestrained structures. For the tightly restrained structure, the values of Δ are less than or equal to the target values of 0·03 Å, 0·04 Å, 0·052 Å and 0·025 Å.

The major difference between the structures obtained from refinements with loose restraints and no restraints is that very large deviations in geometry are absent in the former structure; for example, the r.m.s. deviation from ideality for side-chain angles is 9·9° in the loosely restrained structure and 20·15° in the unrestrained structure.

Apart from the restraints on stereochemistry, the refinement procedure imposes restraints on the absolute differences between the temperature factors of atoms that are bonded together, 1–2 pairs, or bonded to the same third atom, 1–3 pairs (Konnert & Hendrickson, 1980). Based on an analysis of a 30 ps molecular dynamics simulation of bovine pancreatic trypsin inhibitor, Yu *et al.* (1985) support the use of these restraints but indicate that they are about twice as restrictive as they should be.

Let $\Delta = \langle |B_a - B_b| \rangle$, where B_a and B_b are the temperature factors of the two atoms in a 1–2 or 1–3 pair. The average values and standard deviations of Δ for 1–2 and 1–3 pairs between backbone and side-chain atoms are shown in Table 6 for the exact results from the simulation and for the refined structures. The variation of temperature factors is lowest for backbone atoms, as expected, and the exact simulation results show a very large difference between backbone atoms and side-chain atoms. For backbone 1–2 pairs, Δ is 2·87 Å^2, while for side-chain 1–2 pairs Δ is 8·58 Å^2; for 1–3 pairs, the corresponding values are 4·1 and 11·1 Å^2, respectively. The restraint value of Δ is 1·0 Å^2 for backbone and side-chain 1–2 pairs, and 1·5 Å^2 for backbone and side-chain 1–3 pairs (Hendrickson & Konnert, 1980; Hendrickson, 1980). This results in

Table 6
B-*factor variation*

	Backbone bonds $\sigma = 1.0$	Backbone angles $\sigma = 1.5$	Side-chain bonds $\sigma = 1.0$	Side-chain angles $\sigma = 1.5$
Restr1	1·92 (1·59)	2·74 (2·39)	2·73 (2·28)	4·01 (3·63)
Restr2	2·14 (1·81)	2·96 (2·65)	3·03 (2·53)	4·26 (3·91)
Trestr2	0·85 (0·66)	1·39 (1·10)	1·08 (0·86)	1·78 (1·55)
Unrestr1	2·82 (2·91)	3·53 (3·75)	4·46 (4·97)	5·46 (5·69)
Unrestr2	2·78 (2·64)	3·68 (3·52)	4·44 (4·90)	5·59 (5·99)
Altconf	2·03 (1·70)	2·85 (2·52)	2·62 (2·10)	3·83 (3·18)
M.D. (25)	2·87 (4·51)	4·10 (5·94)	8·58 (13·30)	11·11 (18·47)

The refinement program restrains the variation of *B*-factors between 1–2 and 1–3 pairs of atoms in bonds and angles. This Table gives the average and standard deviation (in parentheses) of $|B_i - B_j|$, where i and j are atoms in a 1–2 (bond) pair or 1–3 (angle) pair. These values are to be compared with the target value, σ, for each class of restraint. All values are in *B*-factor units (Å^2). The abbreviations used for the various structures are explained in Table 7. The *B*-factor variations were not restrained in the 2 unrestrained refinements (Unrestr1 and Unrestr2); they are given here for comparison only.

the restrained refinements having values of Δ that are much lower than the exact results (see Table 6). For the unrestrained refinements, the backbone values of Δ are relatively close to the exact results. However, for side-chain pairs, the refined values of Δ are only half as much as the exact values, even though no restraints were placed on them. As we show below, this is consistent with a general trend that is seen in the values of the fluctuations obtained from the refinements.

(c) *Errors in atomic positions*

Root-mean-square deviations between all the refined structures and the dynamical average structure, as well as deviations between the refined structures themselves, are shown in Table 7. The overall r.m.s. error in atomic positions ranges from 0·24 Å to 0·29 Å in the various structures. The errors in backbone positions (0·10 to 0·20 Å r.m.s.) are less than for side-chain atoms (0·28 to 0·33 Å r.m.s.). The backbone errors, though small, are comparable to the r.m.s. deviation of 0·21 Å between the positions of the backbone atoms in the refined experimental structures of oxy- and CO-myoglobin (Phillips, 1980; Kuriyan *et al.*, unpublished results).

The structures from the two loosely restrained refinements are very similar to each other, as are the two structures from the unrestrained refinements. For both backbone and side-chain atoms, the unrestrained refinement results in larger positional errors than the restrained refinement, which is surprising because one might have expected that imposing stereochemical restraints would move the atoms further away from the dynamical average than the unrestrained refinement; the origin of this effect is discussed below. Some backbone atoms in the latter refinement are in error by as much as 0·5 Å, and some side-chain atoms in both refinements are in error by as much as 1·0 Å.

The shifts in atomic positions introduced by tightening the restraints in the refinement (0·09 Å r.m.s. for the backbone and 0·13 Å r.m.s. for the side-chains) is seen to be smaller than the differences between any of the other refined structures and the average dynamical structure, and is comparable to the differences between the two loosely restrained structures. Most importantly, the refinement with tight restraints actually results in the lowest r.m.s. error in backbone positions (0·10 Å). The errors in side-chain positions for this structure are comparable to those in the loosely restrained refinements (0·31 Å) and slightly less than the errors in the unrestrained refinements. Refinement with tight restraints results in a structure having, as expected, stereochemical parameters very much closer to their ideal values than for any other refined structure; yet this structure is as good as or better than structures obtained from refinements with loose restraints or no restraints. The correct average structure is not obtained in any of the refinements.

The positional errors are not uniform over the whole structure. These are plotted as a function

Table 7

Positional deviations between structures

	Restr1	Restr2	Trestr2	Altconf	Unrestr1	Unrestr2
M.D. (25)	0·242	0·260	0·258	0·238	0·286	0·285
	0·132	0·137	0·103	0·145	0·181	0·200
	0·287	0·308	0·314	0·277	0·332	0·320
Restr1		0·116	0·122	0·149	0·197	0·222
		0·063	0·086	0·064	0·114	0·151
		0·140	0·139	0·180	0·232	0·254
Restr2			0·119	0·167	0·214	0·221
			0·087	0·061	0·119	0·150
			0·133	0·204	0·253	0·253
Trestr2				0·178	0·260	0·251
				0·100	0·185	0·159
				0·211	0·294	0·291
Altconf					0·217	0·218
					0·107	0·142
					0·260	0·252
Unrestr1						0·188
						0·126
						0·216

This Table compares the positions of atoms in the 6 refined structures and the molecular dynamics average structure. The backbone atoms (C_α, C, N) of the structures being compared were superimposed by least-squares before the deviations were calculated. There are 3 entries in each box: the 1st entry is the r.m.s. deviation (in Å) between all the atoms in the 2 structures, the 2nd is for just the backbone (C_α, C, N) atoms, and the last entry is for all atoms that are not backbone atoms. The abbreviations for the structures used are explained in Table 3.

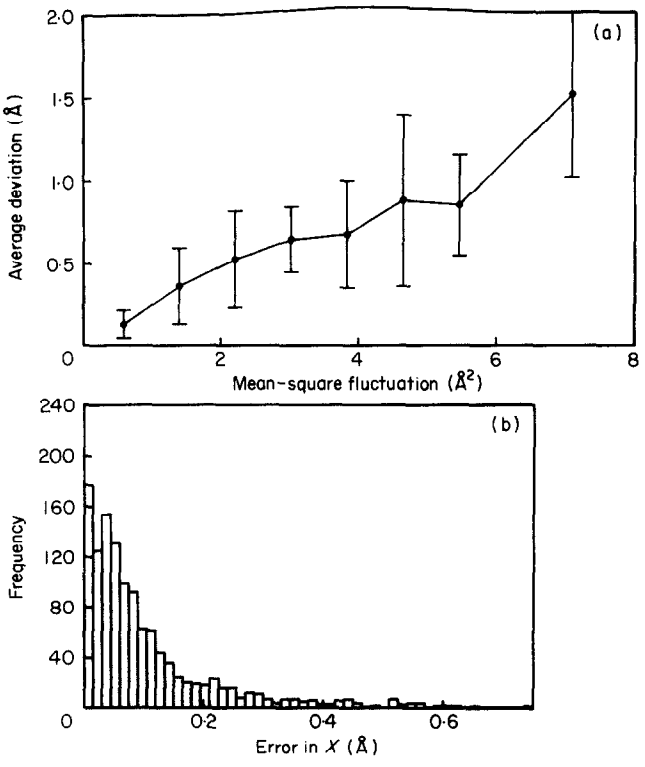

Figure 2. (a) Average positional error as a function of the mean square atomic fluctuations for a restrained refinement of 25 ps data ($\langle F\rangle^{0\cdot05}_{restr}$). The error bars represent ± 1 standard deviation, but for points beyond 2·0 Å^2 there are relatively few points per average. (b) Distribution of positional errors along the principal X-axes for a restrained refinement of the 25 ps data, ($\langle F\rangle^{0\cdot05}_{restr}$).

of residue number for a loosely restrained refinement ($\langle F\rangle^{0\cdot05}_{restr}$) and an unrestrained refinement ($\langle F\rangle^{0\cdot05}_{unrestr}$) in Figure 1. There is a strong correlation between positional error and the magnitude of the mean-square fluctuation for an atom, with certain regions of the protein, such as loops and external side-chains, having greater errors in refined position. We define a local principal axis co-ordinate system for each atom, which is the co-ordinate frame in which the fluctuation second moment tensor is diagonal (Willis & Pryor, 1975). The principal X-axis is taken as the axis along which the fluctuations are the largest, and the Z-axis is taken as that along which the fluctuations are the smallest. In Figure 2(a) the dependence of the positional errors on mean-square fluctuation is shown. Figure 2(b) gives the distribution of positional errors along the principal X-axis for the $\langle F\rangle^{0\cdot05}_{restr}$ structure. The errors in position are largest along the principal X-axis, as expected, as this is the direction of greatest motion.

The co-ordinates obtained by refinement against the 300 ps data have errors that are about twice as large as for the refinements against the 25 ps data, with backbone r.m.s. errors of 0·29 Å and side-chain r.m.s. errors of 0·56 Å. As mentioned above, the errors increase with mean-square fluctuation, and the larger errors are consistent with the larger fluctuations in the 300 ps simulation (see Table 5).

(d) *Errors in refined fluctuations*

The refined mean-square fluctuations are systematically smaller than the fluctuations calculated directly from the simulation for all four refinements. Scatter plots of the fluctuations for all atoms (Fig. 3) show that fluctuations greater than $\sim$0·75 Å^2 ($B = 20$ Å^2) are almost always underestimated by the refinement. Fluctuations less than 0·75 Å^2 are still underestimated more often than they are overestimated in the restrained refinements, though this is less true for the unrestrained refinements. This Figure makes clear the fact that the B-factors (mean-square fluctuations) obtained from the refinement have an effective upper limit independent of the actual values calculated from the dynamics. Figure 4 shows the distribution of errors in the mean-square fluctuations for restrained and unrestrained refinement, and Table 8 gives the correlation coefficients, average absolute errors, average fractional errors and average errors between the four refined sets of temperature factors and the values calculated from the dynamics. The absolute errors averaged over all the atoms range from 5·5 to 7·0 B-factor units (0·21 to 0·27 Å^2), and the average errors range from 4·7 to 7·0 B-factor units (0·18 to 0·27 Å^2). That the average errors are so close to the average absolute errors is in accord with the fact that the refined temperature factors are lower than the values calculated directly from the simulation. The magnitudes and variation of temperature factors along the backbone are reproduced very well by the refinement (Fig. 5(a)) but the refined side-chain fluctuations are almost always too low (Fig. 5(b) and (c)). Regions of the protein that have high mobility also have large errors in refined position and temperature factor.

The average backbone, side-chain and overall B-factors for the various refined structures are compared with the exact simulation results and with experimental B-factors for various liganded forms of myoglobin in Table 5. In the case of backbone temperature factors, both loosely restrained and unrestrained refinements result in a slight lowering of the B-factors, from 12·4 Å^2 to 11·3 Å^2 and 11·7 Å^2, respectively. The damping of the B-factors is much more marked in the side-chain average, which drops from 26·8 Å^2 (exact) to 16·5 Å^2 and 17·6 Å^2 (restrained and unrestrained, respectively). The effect of tightening the restraints is to decrease the variation in the B-factors (see Table 6). This results in the backbone average being slightly larger (12·5 Å^2) than the exact result, and the side-chain average being lowered even further than in the other refinements (14·5 Å^2). The tight restraints result also in the largest errors in B-factors for any of the refinements, with an average fractional error of 36% and an average absolute error of 7·20 B-factor units (0·27 Å^2). This

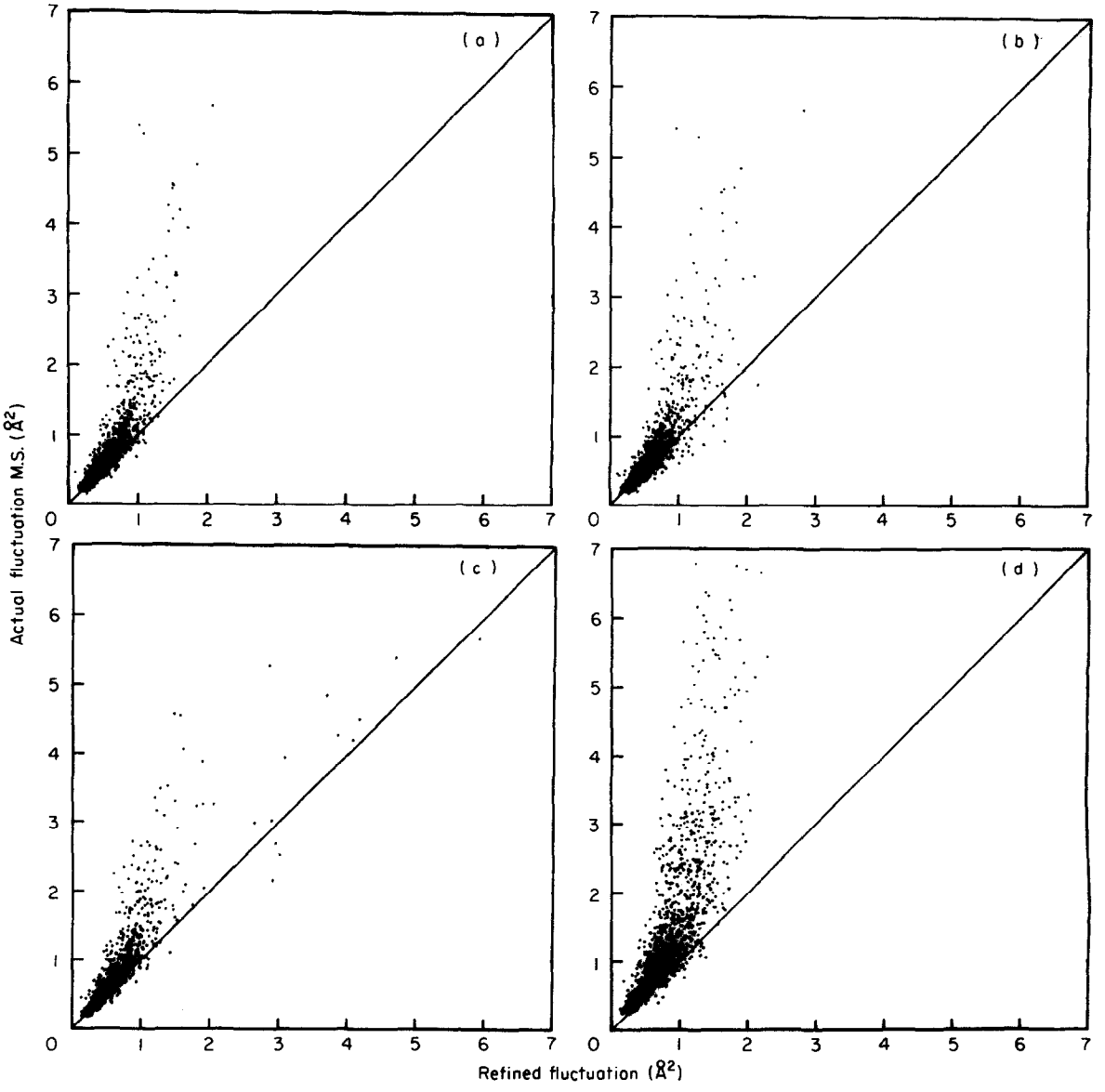

Figure 3. Scatter plots of mean-square (M.S.) fluctuations calculated from the simulation and from the refinements. All the atoms are included in these plots. The exact mean-square fluctuations, $\langle \Delta r_j \rangle^2$, calculated directly from the simulations, are plotted along the y-axis. The refined mean-square fluctuations, obtained from the refined temperature factors are plotted along the x-axis. (a) Results of a restrained refinement ($\langle F \rangle^{0.05}_{\text{restr}}$) of the 25 ps data. (b) Results of an unrestrained refinement ($\langle F \rangle^{0.05}_{\text{unrestr}}$) of the 25 ps data. (c) Results of refinement of the 25 ps data with 10 residues modelled with disordered side-chains ($\langle F \rangle^{0.05}_{\text{altconf}}$). The refined mean-square fluctuations for atoms with more than 1 conformation were obtained by averaging over both conformations. (d) Refinement of the 300 ps data with loose restraints. These results were obtained by re-refining the model obtained from the initial refinement after increasing all the B-factors by 13·5 Å^2 and by scaling the B-factors to minimize the R-factor (see Fig. 6 and the text).

is in contrast to the positional errors, where the tight restraints resulted in the lowest errors.

The systematic underestimation of the B-factors by the refinement increases as the fluctuations increase (see Fig. 3), and so one might expect that the B-factors in the 300 ps simulation would be greatly reduced on refinement, since the average fluctuations are twice as large over this time-scale than over a 25 ps time-scale (Levy *et al.*, 1985). This is indeed seen to be the case, and Table 5 gives the average B-factors for the exact simulation results and the refinement. The backbone and side-chain B-factors drop from 25·6 Å^2 and 48·6 Å^2 to 16·8 Å^2 and 21·1 Å^2, respectively.

To check for a systematic error in the B-factor (a constant offset), R-factors were calculated as a function of a constant shift applied to the refined B-factors. The result, for refinement of the 300 ps data, is shown in Figure 6, which shows that there is no constant offset to the B-factors that would improve the R-factor. This was the case also for all the 25 ps refinements. The 300 ps refinement had been started with an initial B-factor of 15·0 Å^2, which is much lower than the exact average

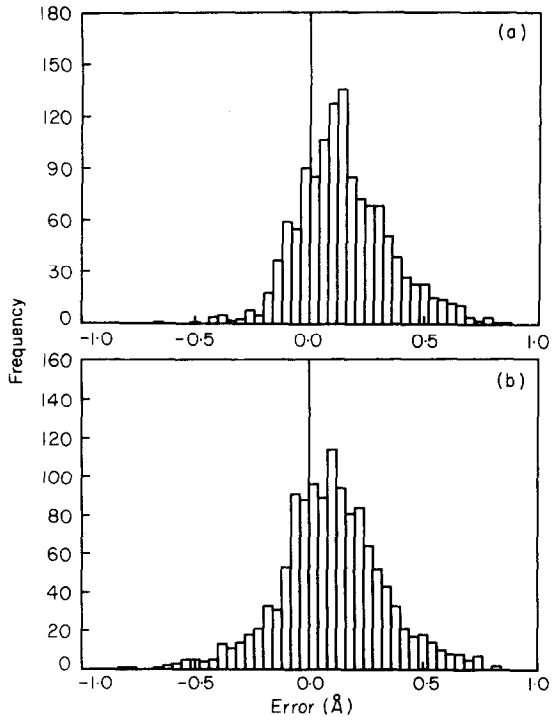

Figure 4. Distribution of errors in mean-square fluctuation for a restrained and an unrestrained refinement of the 25 ps data. The error is defined as $\sigma^2_{md}-\sigma^2_{ref}$ (see the text). (a) Distribution of errors in the restrained refinement results. (b) Distribution of errors in the unrestrained refinement results.

B-factor (Table 5). The effect of the initial *B*-factor on the final values of the refined *B*-factors was tested by increasing all the *B*-factors in the refined model by 13·5 *B*-factor units (0·5 Å^2) and continuing the refinement. The final *B*-factors obtained in this way were higher than the previous results, with backbone and side-chain averages of 22·2 Å^2 and 26·5 Å^2, respectively; an increase of about 3·0 Å^2. Once again we checked whether a constant *B*-factor offset would improve the *R*-factors, and the *R*-factor as a function of *B*-factor shift is shown in Figure 6. This time it is seen that a *B*-factor shift of −3·0 Å^2 reduces the *R*-factor by about 1·5%. This shift is in accord with the original *B*-factor values. This *B*-factor shift was applied to all the atomic temperature factors and Figure 3(d) shows a scatter plot of these final mean-square fluctuations against the exact results from the simulation. The underestimation of the fluctuations is a striking feature of this plot.

(e) *The anisotropy and anharmonicity of the atomic fluctuations*

(i) *The anisotropy*

In this and the following sections, some of the characteristics of the atomic probability distribution functions obtained from the simulation will be discussed. The probability distribution function, $p(\mathbf{u})$, gives the probability density of finding the atom displaced **u** from its mean position. Probability distribution functions are characterized by their mean, **m**, and their higher moments; the second moment, σ^2, is the most important in crystallography, since it is related to the temperature factor. The average electron density of an atom is the convolution of its electron density at rest with its probability distribution function. Since the atomic electron density can be considered to be time-independent, the probability distribution function can be used instead of the average electron density (Willis & Pryor, 1975). We shall use σ^2_{md} to mean the exact mean-square fluctuation calculated from the simulation, and σ^2_{ref} to mean that obtained from refinement.

The anisotropy and anharmonicity of the distribution functions in molecular dynamics simulations of proteins have been studied in detail (Mao *et al.*, 1982; Ichiye & Karplus, unpublished results). Rather than carrying out an extensive analysis of the probability distributions in the simulation used here, it will be shown that the magnitudes of the anisotropy and anharmonicity in myoglobin are very similar to those found in the 30 ps simulation of lysozyme by Ichiye & Karplus (unpublished results). The approach of these authors is used in what follows. Let U_x, U_y and U_z be the fluctuations from the mean position along the principal X, Y and Z axes and:

$$\sigma^2_x = \langle U^2_x \rangle,\ \sigma^2_y = \langle U^2_y \rangle \text{ and } \sigma^2_z = \langle U^2_z \rangle, \quad (15)$$

with $\sigma_x \geq \sigma_y \geq \sigma_z$. We define one measure of the anisotropy by:

$$A_1 = \left[\frac{\sigma^2_x}{\frac{1}{2}(\sigma^2_y+\sigma^2_z)}\right]^{1/2} - 1{\cdot}0. \quad (16)$$

This measures the amount by which the ratio of the fluctuations in the principal X-direction to the average of that in the other two directions exceeds that of an isotropic distribution, for which A_1 is zero. We also define another measure of the anisotropy:

$$A_2 = \left[\frac{\sigma^2_y}{\frac{1}{2}(\sigma^2_y+\sigma^2_z)}\right]^{1/2} - 1{\cdot}0, \quad (17)$$

which measures how isotropic the motion is in the principal Y–Z plane. A_1 and A_2 have been calculated for various classes of atoms and the values are given in Table 9, which includes the (unpublished) results of Ichiye & Karplus for the same classes of atoms in lysozyme. While the anisotropy defined either way is seen to be slightly lower in myoglobin, the general trends are the same in both molecules. The motions tend to be quite anisotropic. Very few atoms (about 1·4%) have A_1 less than 0·02; 61% of the atoms have A_1 greater than 0·5, and 31% have A_1 greater than 0·75. Atoms further out along the side-chain have higher values of A_1, but the value of A_2 remains uniformly low for all classes of atoms (at about 0·15). This indicates that the most significant contribution to the anisotropy is along the direction of largest

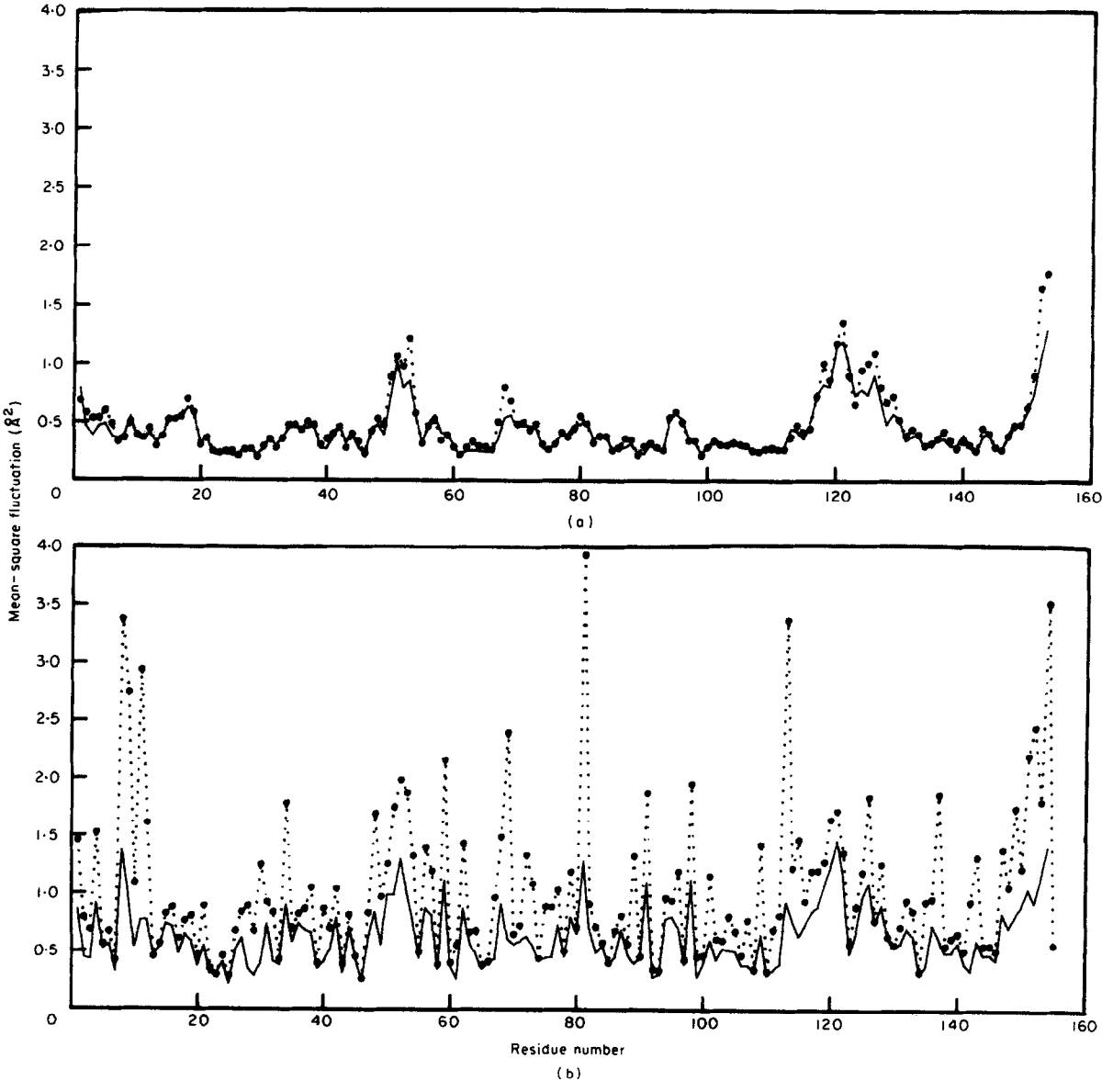

Figure 5. Residue averages of mean-square fluctuations from molecular dynamics (dotted line) and refinements (continuous line). All plots are for the results of refining the 25 ps data. (a) Backbone (N, C and C_α) averages for the restrained $\langle F\rangle^{0\cdot05}_{restr}$ refinement. (b) Side-chain averages for a restrained ($\langle F\rangle^{0\cdot05}_{unrestr}$) refinement. (c) Side-chain averages for an unrestrained ($\langle F\rangle^{0\cdot05}_{unrestr}$) refinement. (d) Side-chain averages for the refinement of the ($\langle F\rangle^{0\cdot05}_{unrestr}$) data with alternate conformations for 10 side-chains.

motion, and that the motion is rather more isotropic along the principal Y–Z plane.

Ichiye & Karplus (unpublished results) have studied the errors introduced by refining an isotropic Gaussian model for the distribution function against analytic probability distributions and distributions obtained from the simulation. The procedure they use is analogous to real space refinement of the electron density of an isolated atom; instead of refining a model for the electron density, they refine models for the distribution function. The function they minimize is of the form:

$$R = \int (p_c(\mathbf{u}) - p_o(\mathbf{u}))^2 \, d\mathbf{u}, \qquad (18)$$

where p_o and p_c are the actual and model distributions, respectively. Diamond (1971) has shown that this kind of refinement is equivalent to reciprocal space refinement where all the structure factors are weighted equally. In the refinement procedures we use here, the structure factors are not weighted equally (eqn (14)). Nevertheless, it is helpful to use the results for lysozyme to aid in the analysis of the reciprocal space refinements reported here.

To examine the effect of anisotropy separately from that of anharmonicity, Ichiye & Karplus (unpublished results) studied the case where p_o is a three-dimensional anisotropic Gaussian and p_c is an isotropic Gaussian. They found that the refined values of σ are close to the actual values only for small values of the anisotropy ($\sigma_x/\sigma_y < 1\cdot5$). For larger anisotropies, the refined values of σ are

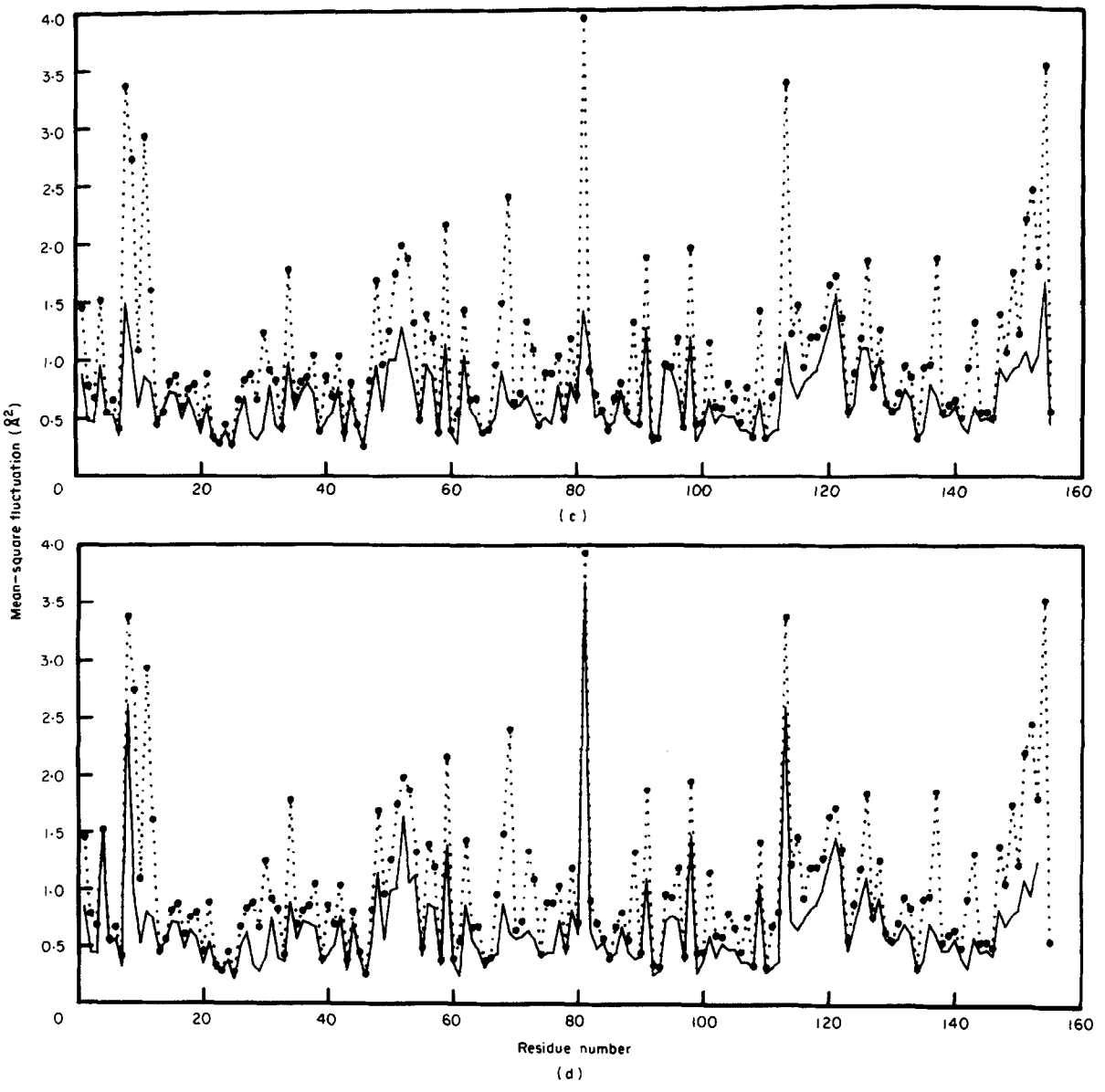

Fig. 5.

always lower than the actual value. This suggests that there should be a correlation between the values of the anisotropy in the myoglobin simulation with the errors in the refined values of the temperature factor. Figure 7 shows the dependence of the position and fluctuation errors on anisotropy for one of the refinements with loose restraints. The errors in position and temperature factor increase with anisotropy, but the effects of anisotropy cannot be separated from those of anharmonicity. Also, if the probability distribution function were an anisotropic Gaussian, there would be no predicted error in the refined position; however, the error in position also increases with anisotropy.

(ii) *The anharmonicity*

The third and fourth moments of the distribution can be used to characterize the anharmonicity (Mao *et al.*, 1982; Ichiye & Karplus, unpublished results). The skewness, α_{3i}, where i is x, y or z, is defined by:

$$\alpha_{3i} = \frac{\langle U_i^3 \rangle}{\langle U_i^2 \rangle^{3/2}}, \tag{19}$$

and the coefficient of excess kurtosis, α_{4i}, is given by:

$$\alpha_{4i} = \frac{\langle U_i^4 \rangle}{\langle U_i^2 \rangle^2} - 3{\cdot}0. \tag{20}$$

Both α_3 and α_4 are zero for a Gaussian distribution. The average values of $|\alpha_{3i}|$ and $|\alpha_{4i}|$ for various classes of atoms have been calculated and compared with the values obtained for lysozyme (Ichiye & Karplus, unpublished results). The values for the two proteins are strikingly similar (Table 10). From a detailed study of the moments of the atomic distributions, Ichiye & Karplus conclude that most atoms with large anharmonicity have multiple

Table 8
Comparison of B*-factors*

	Restr1	Restr2	Trestr2	Altconf	Unrestr1	Unrestr2
M.D. (25)	0·82	0·82	0·69	0·90	0·80	0·80
	30·7	30·4	35·9	25·1	35·1	29·2
	6·16	6·09	7·20	5·04	7·03	5·84
	5·71	5·61	6·27	4·53	6·58	4·72
Restr1		0·99	0·93	0·78	0·94	0·95
		5·2	14·2	12·1	14·1	12·5
		0·75	2·03	1·73	2·02	1·78
		−0·10	0·56	−1·18	0·87	−0·99
Restr2			0·92	0·78	0·94	0·96
			15·1	11·4	14·1	10·7
			2·18	1·64	2·03	1·54
			0·66	−1·08	0·97	−0·89
Trestr3				0·66	0·82	0·84
				23·7	22·9	22·2
				3·26	3·15	3·05
				−1·74	0·31	−1·55
Altconf					0·76	0·76
					19·3	16·2
					2·99	2·51
					2·05	0·19
Unrestr1						0·96
						17·4
						2·35
						−1·86

This Table compares the B-factors of the 6 refined structures and the exact B-factors obtained from the 25 ps simulation. The abbreviations used for the structures are the same as in Table 7. There are 4 entries in each box:

First entry: the correlation coefficient between the B-factors for all atoms in 2 structures being compared.

Second entry: the percentage difference defined as:

$$\text{percentage difference} = \frac{\sum |B_i - B_j|}{\sum B_i} \times 100{\cdot}0,$$

where B_i refers to the ith row and B_j to the jth column.

Third entry: the average absolute error $|B_i - B_j|$.

Fourth entry: the average error $B_i - B_j$.

peaks in their distribution functions, with each peak being close to harmonic. They suggest that the best description of anharmonicity for atoms with large fluctuations should not be based on perturbations to a local Gaussian distribution, but should include contributions from separated Gaussian distributions.

(f) *Probability distribution functions from dynamics and refinement*

It is of interest to determine whether the errors in the refinement are localized to just a few residues of the protein, and to determine what causes the systematic underestimation of the fluctuations. In Table 11, all the residues that have at least one atom with a refined temperature factor 50% lower than the exact value are listed. The temperature factors used are from the refinement of the $\langle F \rangle^{0{\cdot}05}$ data with loose restraints. There are 77 such atoms distributed over 45 residues, and Table 11 includes the secondary structure elements (helices or loops) and average solvent-accessible area for these residues.

Most of the residues are in the helix regions, with only seven in the inter-helix loops. Among the helices, the B and F helices have the lowest number of such residues, with two each, and the A helix has the most, with ten. Most of the residues have charged side-chains and are on the surface, but 11 of them are partially or completely buried (with average side-chain solvent-accessibility less than 3·0 Å^2). These include a tryptophan, a valine, four leucine, an isoleucine, a glycine, a histidine, an arginine and an aspartic acid residue.

The probability distribution functions from the 25 ps simulation of met-myoglobin for about 30 such atoms have been studied in the following way. For each atom, the second moment tensor was calculated and diagonalized to obtain the transformation to the local principal axis frame. The time-series for the three principal axis co-ordinates were calculated from the simulation and, from the time-series, the probability of fluctuations along the three principal axes was estimated by dividing the co-ordinate ranges into 25 bins and counting the number of times the trajectory was in each bin. The resulting distribution was normalized to have

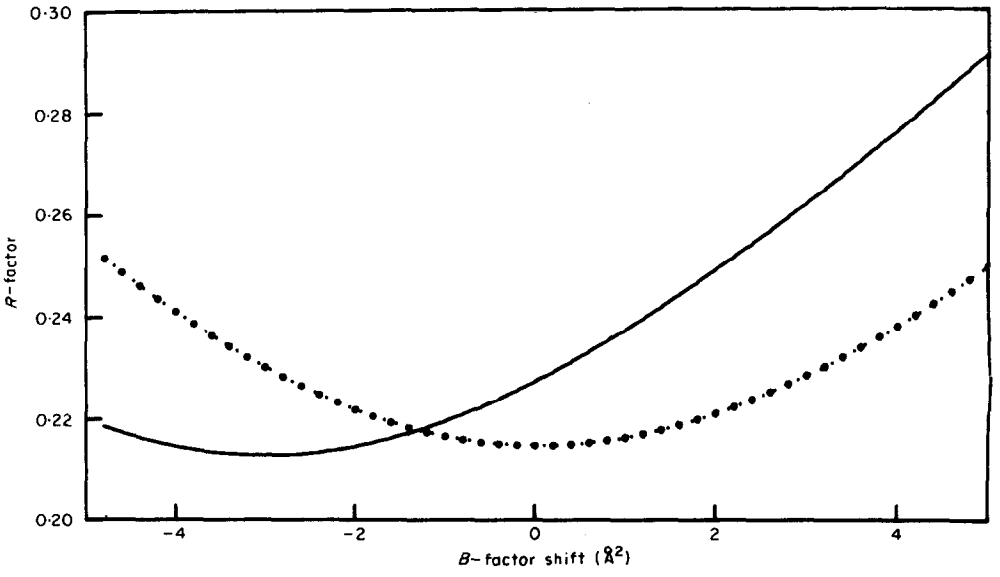

Figure 6. *R*-factor *versus* ΔB. The *R*-factor as a function of a uniform shift in temperature factor, ΔB, is plotted for the initial refined structure (300 ps data; dotted line) and the structure obtained from increasing the *B*-factors by 13·5 *B*-factor units and re-refining (continuous line). The *R*-factor is given by:

$$R = \frac{\sum |F_o - F'_c|}{\sum |F_o|},$$

where F'_c is given by:

$$F'_c(\mathbf{Q}) = F_c(\mathbf{Q})\, e^{-(\Delta B)s^2}.$$

F_c is the structure factor calculated from the refined model.

unit total probability. For each atom, the Gaussian distributions corresponding to the simulation average position and σ^2_{md} and the refined position and σ^2_{ref} were calculated. For every such atom studied, the simulation had two or more well-separated regions of high probability and the refinement had fit only one of the regions, neglecting the rest. This explains both the lower refined fluctuation and the error in the refined position, as the refinement moves the atom from the true average position into one of the regions of high probability.

Figure 8(a) and (b) shows the molecular dynamics distribution function along the principal *X*-axis as well as the equivalent Gaussian distribution and the refined (restrained and unrestrained) Gaussian distributions for the $C_{\varepsilon 1}$ atom in histidine 81. The exact distribution function has two major peaks, and both restrained and unrestrained refinements fit only one of the two peaks. The positional error in the unrestrained refinement is larger because it has moved even further towards fitting just one peak. Two more examples of this kind of distribution are shown in Figure 8. They are for the $C_{\delta 1}$ atom of leucine 69 (Fig. 8(d)) and for the $C_{\delta 2}$ atom of leucine 11 (Fig. 8(e)). The former has two well-separated peaks and a long tail in the distribution, and the refinement fits only the major peak. The distribution for leucine 11 is interesting because it has three peaks, and the refinement fits two of them but not the third.

These results are consistent with the findings of Ichiye & Karplus (unpublished results). On refining molecular dynamics distributions with isotropic Gaussians, they found that the refinement would usually fit the major peak of a multi-peaked distribution and neglect the rest. They also studied the refinement of a double-peaked Gaussian distribution (2 identical isotropic Gaussians separated along the principal *X*-axis) by a single isotropic Gaussian. The solution to this problem is obtained analytically. For small values of the separation, δ, between the two peaks, the refinement will fit both Gaussians. However, when

Table 9
Statistics on anisotropy A_1

	Myoglobin	Lysozyme
All atoms	0·68 (0·39)	0·85 (0·55)
Backbone	0·57 (0·28)	0·77 (0·50)
Side-chain	0·74 (0·43)	0·93 (0·59)
N	0·55 (0·26)	0·68 (0·30)
C	0·58 (0·28)	0·76 (0·45)
O	0·70 (0·40)	0·93 (0·60)
C_α	0·59 (0·30)	0·73 (0·47)
C_β	0·67 (0·40)	0·74 (0·45)
γ	0·72 (0·46)	0·90 (0·55)
δ	0·76 (0·45)	0·95 (0·56)
ε	0·85 (0·47)	1·03 (0·67)
ζ	0·77 (0·42)	1·14 (0·73)

The numbers are averages over all atoms for a particular class, except that proline residues were excluded for outer side-chain averages. Numbers in parentheses are standard deviations. The results for lysozyme are taken from Ichiye & Karplus (unpublished results). The results for Mb are from the 25 ps simulation sampled every 0·05 ps. The value of A_2 is the same (approx. 0·15) for all classes of atoms.

Table 10

Statistics on skewness and kurtosis by atom-type for myoglobin and lysozyme

A. *Skewness* $|\alpha_3|$ *by atom type for myoglobin*

	U_x	U_y	U_z
All atoms	0·38 (0·32)	0·28 (0·25)	0·21 (0·21)
Backbone	0·36 (0·28)	0·26 (0·24)	0·21 (0·17)
Side-chain	0·40 (0·34)	0·29 (0·26)	0·22 (0·24)
N	0·36 (0·27)	0·25 (0·22)	0·20 (0·18)
C	0·37 (0·28)	0·26 (0·25)	0·22 (0·17)
O	0·41 (0·34)	0·26 (0·20)	0·22 (0·17)
C_α	0·35 (0·28)	0·27 (0·26)	0·21 (0·17)
C_β	0·34 (0·30)	0·33 (0·28)	0·20 (0·16)
γ	0·40 (0·37)	0·30 (0·25)	0·23 (0·42)
δ	0·38 (0·34)	0·26 (0·25)	0·21 (0·18)
ε	0·40 (0·35)	0·33 (0·31)	0·24 (0·20)
ζ	0·43 (0·34)	0·31 (0·30)	0·22 (0·16)

Numbers are averages over all the atoms of a particular class, except that outer side-chain averages were omitted for proline residues. Numbers in parentheses are standard deviations. These results are for met-myoglobin (25 ps).

B. *Skewness* $|\alpha_3|$ *by atom type for lysozyme*

	U_x	U_y	U_z
All	0·38 (0·32)	0·25 (0·23)	0·18 (0·16)
Backbone	0·34 (0·28)	0·22 (0·20)	0·17 (0·14)
Side-chain	0·42 (0·36)	0·28 (0·26)	0·20 (0·18)
N	0·30 (0·25)	0·21 (0·16)	0·16 (0·11)
C	0·33 (0·27)	0·22 (0·21)	0·17 (0·13)
O	0·38 (0·33)	0·27 (0·25)	0·18 (0·15)
C_α	0·33 (0·25)	0·19 (0·17)	0·18 (0·15)
C_β	0·32 (0·24)	0·24 (0·22)	0·17 (0·14)
γ	0·40 (0·36)	0·25 (0·20)	0·18 (0·15)
δ	0·45 (0·38)	0·31 (0·30)	0·21 (0·22)
ε	0·53 (0·51)	0·32 (0·27)	0·22 (0·20)
ζ	0·47 (0·36)	0·30 (0·27)	0·21 (0·15)

These data are taken from Ichiye & Karplus (unpublished results), and are from a 30 ps simulation of lysozyme.

C. *Kurtosis* $|\alpha_4|$ *by atom-type for myoglobin*

	U_x	U_y	U_z
All atoms	0·58 (0·58)	0·45 (0·46)	0·36 (0·67)
Backbone	0·56 (0·46)	0·43 (0·36)	0·36 (0·33)
Side-chain	0·59 (0·64)	0·46 (0·51)	0·37 (0·80)
N	0·53 (0·51)	0·42 (0·33)	0·36 (0·44)
C	0·55 (0·42)	0·45 (0·38)	0·34 (0·25)
O	0·56 (0·67)	0·42 (0·35)	0·31 (0·29)
C_α	0·58 (0·44)	0·44 (0·36)	0·37 (0·27)
C_β	0·49 (0·46)	0·52 (0·61)	0·36 (0·47)
γ	0·60 (0·75)	0·43 (0·44)	0·48 (1·75)
δ	0·63 (0·56)	0·46 (0·48)	0·35 (0·28)
ε	0·68 (0·71)	0·52 (0·69)	0·34 (0·27)
ζ	0·72 (0·72)	0·53 (0·53)	0·33 (0·34)

Numbers in parentheses are standard deviations. The data are for met-myoglobin and are calculated from the 25 ps simulation.

D. *Kurtosis* $|\alpha_4|$ *by atom-type for lysozyme*

	U_x	U_y	U_z
All	0·56 (0·52)	0·39 (0·49)	0·31 (0·36)
Backbone	0·50 (0·43)	0·33 (0·37)	0·27 (0·25)
Side-chain	0·61 (0·59)	0·46 (0·58)	0·35 (0·44)
N	0·46 (0·38)	0·30 (0·24)	0·24 (0·19)
C	0·48 (0·38)	0·33 (0·33)	0·26 (0·22)
O	0·57 (0·56)	0·37 (0·56)	0·27 (0·21)
C_α	0·50 (0·38)	0·32 (0·27)	0·31 (0·36)
C_β	0·48 (0·37)	0·36 (0·36)	0·28 (0·20)
γ	0·57 (0·48)	0·42 (0·39)	0·26 (0·25)
δ	0·65 (0·53)	0·54 (0·79)	0·45 (0·72)
ε	0·85 (1·07)	0·61 (0·72)	0·46 (0·51)
ζ	0·67 (0·52)	0·40 (0·52)	0·41 (0·24)

$\delta > 3{\cdot}74\sigma_0$, where σ_0^2 is the second moment of one of the two Gaussians in the double-peaked distribution, the refinement will fit only one of the Gaussians with a refined σ only slightly larger than σ_0.

The distribution functions for atoms that have low fluctuations in the simulation but have high refined temperature factors have been examined. Such atoms are usually close to atoms that have large mean-square fluctuations and multiple peaks in the distribution, and the larger refined fluctuation was seen to be due to the refinement moving these atoms towards the extra density of the disordered atoms. This feature was most marked in the unrestrained refinement, and the C_γ atom of leucine 11 is an example. The terminal atoms of this residue, $C_{\delta 1}$ and $C_{\delta 2}$, are disordered (see above), and the exact temperature factors for the C_γ, $C_{\delta 1}$ and $C_{\delta 2}$ atoms are 19·46 Å^2, 63·5 Å^2 and 71·4 Å^2, respectively. The fluctuations obtained from unrestrained refinement are 31·9 Å^2, 26·9 Å^2 and 24·3 Å^2, respectively, with large errors in the refined position of all three atoms, and with the temperature factor of the C_γ atom significantly overestimated. Examination of the distribution functions for this atom along with the distribution functions for the terminal atoms clearly showed that C_γ atom moves to fit the extra density due to the disordered terminal atoms.

Refinement with restraints prevents atoms from moving into the density of neighbouring atoms, as this would lead to violations of stereochemistry. The result is a lowering of the errors in well-defined atoms that are close to disordered atoms, and this explains the fact that refinements with restraints yield structures closer to the true average.

The large separation between the peaks in some of the distributions examined suggested that several residues probably needed to be modelled by including alternate conformations. The exact distributions were not used to decide which residues to model in this way, because this information is never accessible in an experimental situation. Instead, difference electron density maps were calculated. As

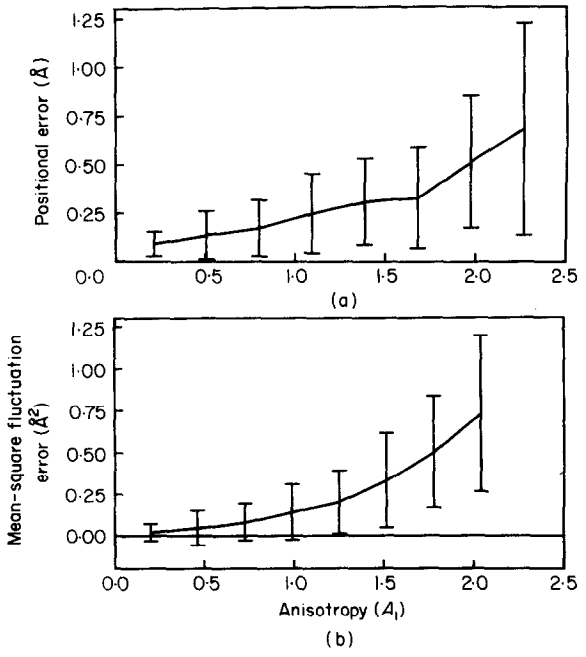

Figure 7. Errors in the positions and fluctuations *versus* the anisotropy A_1 for restrained refinement of the 25 ps data. The errors are averaged in bins and the values are plotted with ± 1 standard deviation bars. All errors are calculated from the refinement of $\langle F \rangle^{0.05}$ data with loose restraints. (a) Positional error; (b) error in fluctuations, where error is given by $\sigma^2_{md} - \sigma^2_{ref}$.

described in Methods, a $(2F_o - F_c)\exp(i\alpha_c)$ synthesis was used, where the phases, α_c, and the model structure factors, F_c, were calculated from the $\langle F \rangle^{0.05}_{restr}$ structure. F_o is the amplitude of the structure factor calculated from the simulation.

For ten residues, the difference map clearly showed the existence of alternate side-chain conformations that were not accounted for in the refined model. These alternate conformations were modelled by changing the side-chain torsion angles and fitting the extra electron density. For other residues, the situation was not so clear, and building in alternate conformations would have required some judgement. This is partly because the errors in the refined model used to phase the data affect the quality of the difference map (Blundell & Johnson, 1976), and because the limited resolution of the data (1·5 Å) makes it difficult to identify the conformations that are separated by 0·5 Å to 1·5 Å or less. Another difficulty in modelling the disorder is that simple side-chain torsional isomerization may be an inadequate model of the complicated dynamics actually taking place. Some examples are given below of residues that sample multiple conformations without undergoing torsional transitions.

The refinement of a model including alternate conformations for ten side-chains (these are listed in Table 11) is described in Methods. The refinement works very well for residues that had very large errors in the previous refinement. Figure 8(c) shows the distribution functions from molecular dynamics and from the alternate conformations refinement for histidine 81 $C_{\beta 1}$. The double-peaked Gaussian describes the actual molecular dynamics distribution rather well. If the average position and temperature factor for the atom are calculated from its two refined positions and occupancies, we see that the error in refined position is only 0·18 Å, and the error in refined mean-square fluctuation is only 0·25 Å^2. Refinement with only one conformation led to errors of 2·0 Å in position and 5·7 Å^2 in mean-square fluctuation.

However, the overall agreement between the refined fluctuations and the dynamics fluctuations is not greatly improved by including alternate conformations for just ten residues. Figure 3(c) shows a scatter plot of all the fluctuations for the dynamics and the refinement with alternate conformations, and the effect of this refinement is seen to be an improvement in the agreement only for atoms that previously had extremely high errors. The same is true of the positional errors, and Table 6 includes the correlations and errors for this refinement, which are similar to those for all the other refinements. Figure 5(d) shows the fluctuations as a function of residue number for this refinement. It appears that there might be a few side-chains for which alternate conformations might still be built but, for the reasons mentioned above, it is difficult to do much more with a 1·5 Å resolution difference map.

Refining alternate conformations for ten residues has not lowered the *R*-factor very much. The final *R*-factor for restrained refinement is 13·6%, and for restrained refinement with alternate conformations it is 12·6%, indicating that the major inadequacies in the refinement model have not been removed by refining alternate conformations for just a few residues.

(g) *An examination of structural transitions*

The well-separated regions of density seen for residues with large errors implies the existence of structural transitions from one local potential minimum to another. An exhaustive analysis would require the calculation and examination of the positional time-series of all the atoms in question, which has not been attempted. Here, two preliminary analyses are presented, one of dihedral transitions and another of larger-scale helix deformations.

Dihedral transitions were monitored by following trajectories from one minimum in the torsional potential to another for all the dihedral angles in the protein. This analysis was done using the program CHARMM (Brooks *et al.*, 1983). A transition is defined as a change in the dihedral angle from one well of the torsional potential to another, the wells in the potential being defined by the periodicity of the energy function for that torsion (Brooks *et al.*, 1983). A transition is counted as such only if it involves crossing at least 30°

Table 11
Residues with atoms that have large errors in B*-factor*

Helix	Residue number	Residue type	Atom	Torsional transitions	Accessible surface areas (Å^2)	
					Backbone	Side-chain
NA1	1	Val	$C_{\gamma 1}$		11·3	17·2
A1	3	Ser	C_β	None	5·7	18·3
A2	4*	Glu	$O_{\varepsilon 2}$	$\chi_{1,2,3}$	0·0	17·1
A5	7	Trp	$C_{\zeta 3}$	$\chi_{2,3}$	0·0	0·5
A6	8*	Gln	$N_{\varepsilon 2}$	$\chi_{2,3}$	0·7	14·7
A7	9	Leu	$C_{\delta 1}$	$\chi_{1,2}$	2·3	10·1
A8	10	Val	$C_{\gamma 2}$	χ_1	0·0	0·0
A9	11	Leu	$C_{\delta 2}$	$\chi_{1,2}$	0·2	2·9
A10	12	His	$C_{\varepsilon 1}$	None	3·5	14·5
B11	30	Ile	$C_{\gamma 2}$	$\chi_{1,2}$	0·0	1·7
B15	34	Lys	N_ζ	ϕ, $\chi_{1,2,3}$	2·8	13·9
C3	38	Glu	$O_{\varepsilon 1}$	$\chi_{2,3}$	2·3	9·5
CD6	48*	His	$N_{\varepsilon 2}$	ϕ, ψ, χ_1	2·6	14·1
D1	51	Thr	$C_{\gamma 2}$	None	0·2	18·7
D2	52*	Glu	$O_{\varepsilon 2}$	$\chi_{1,2,3}$	0·3	10·8
D3	53	Ala	C_β	ψ	5·2	38·5
D4	54	Glu	$O_{\varepsilon 1}$	ϕ, ψ, $\chi_{1,2,3}$	1·9	8·0
D5	56	Lys	N_ζ	ψ, $\chi_{2,3,4}$	0·1	15·3
E2	59*	Glu	$O_{\varepsilon 2}$	$\chi_{1,2,3}$	0·1	18·1
E5	62	Lys	N_ζ	$\chi_{1,2,3,4}$	0·0	7·9
E10	67	Thr	$C_{\gamma 2}$	ψ	2·1	5·7
E12	69	Leu	$C_{\delta 1}$	ψ, χ_2	0·0	0·0
E16	73	Gly	O	ψ	0·4	–
EF4	81*	His	$N_{\varepsilon 2}$	$\chi_{1,2}$	0·1	13·8
EF5	82	His	$N_{\varepsilon 2}$	None	0·0	0·0
F4	89	Leu	$C_{\delta 2}$	$\chi_{1,2}$	0·0	2·4
F6	91	Gln	$N_{\varepsilon 2}$	$\chi_{1,2,3}$	0·1	10·5
FG2	96	Lys	N_ζ	χ_1	4·1	18·0
FG4	98*	Lys	N_ζ	$\chi_{2,3,4}$	0·0	14·1
G10	109*	Glu	$O_{\varepsilon 1}$	$\chi_{1,2,3}$	0·3	7·5
G14	113*	His	$C_{\varepsilon 1}$	$\chi_{1,2}$	2·5	9·5
G16	115	Leu	$C_{\delta 1}$	χ_2	0·0	0·0
G19	118	Arg	N_{H2}	$\chi_{1,2,3,4}$	2·4	8·9
H2	126	Asp	$O_{\delta 1}$	χ_2	6·2	16·0
H12	136	Glu	$O_{\varepsilon 1}$	$\chi_{1,2,3}$	0·0	7·7
H13	137	Leu	$C_{\delta 1}$	$\chi_{1,2}$	0·0	4·3
H15	139	Arg	N_{H2}	χ_3	0·3	2·7
H17	141	Asp	O	ψ	0·0	0·0
H19	143	Ala	O	ψ	0·0	11·2
H23	147	Lys	N_ζ	$\chi_{1,2,3,4}$	1·5	22·0
HC4	152	Gln	C_γ	ϕ, ψ, χ_3	7·8	6·1

The Table is a list of the residues which, in the 25 ps simulation of met-myoglobin, have at least 1 atom with a refined *B*-factor less than half the exact value. The refined structure used is from the refinement of $\langle F \rangle^{0.05}$ with loose restraints. For each such residue, the torsion angles that undergo transitions are listed. Since the transitions could be transient, this list does not imply that alternate conformations corresponding to different equilibrium values for these torsions are actually sampled to a significant extent in the simulation (see the text). The average accessible surface areas for the backbone and side-chain atoms in the residue are listed; a water-sized spherical probe was used in the surface calculations. The residues that were modelled by 2 alternate conformations in the refinement are labelled with an asterisk.

beyond the maximum of the barrier. Torsional angles that underwent transitions in the 45 residues with large error are listed in Table 11.

The time-series for these torsion angles have not been examined to see if the transitions are merely transient jumps to another well or if they actually represent a significant population of two or more conformations. Nevertheless, they indicate the kinds of motion that might lead to disorder in the side-chain or backbone. Of the side-chain torsions, transitions in χ_1 or χ_2 lead to the largest shifts in side-chain position, and the ten residues that were modelled by alternate conformations in the refinement all have transitions in at least one of these dihedrals. Many residues also have transitions in the backbone ϕ and ψ dihedrals and, for some, this results in the carbonyl oxygen of the backbone being disordered. Figure 8(f) shows the distribution functions for the carbonyl oxygen of aspartate 141, and once again it is seen that the refinement fits only the major peak of a multi-peaked distribution.

Four of the residues in Table 11 actually have no transitions in torsional angles, either for the side-chains or for the backbone. However, they too have

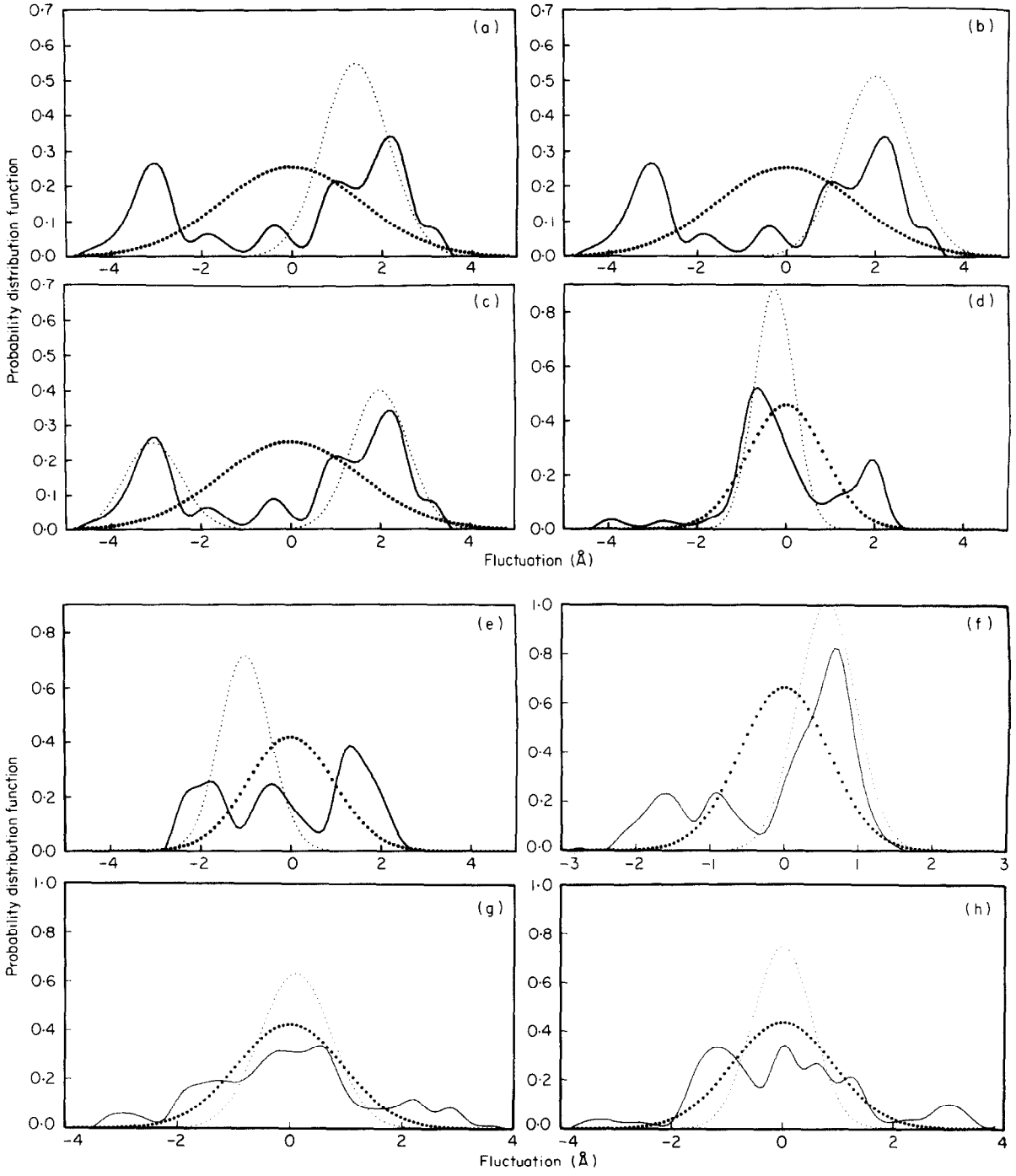

Figure 8. Probability distribution functions along the principal X-axis for the 25 ps data. Continuous line: exact distribution function calculated from the simulation. A spline was used to smooth the data. Large dots, the Gaussian determined by the average molecular dynamics position and $\sigma^2_{md}/3$. Small dots, the Gaussian determined by the refined position and $\sigma^2_{ref}/3$. (a), (b) and (c) are for the $C_{\varepsilon 1}$ atom of histidine 81. (a) Restrained refinement; (b) unrestrained refinement; and (c) refinement with 2 conformations for the histidine side-chain. The other distributions shown are for unrestrained refinements of (d) leucine 69 $C_{\delta 1}$, (e) leucine 11 $C_{\delta 2}$, (f) aspartate 141 O, (g) threonine 51 C_γ and (h) histidine 12 $C_{\varepsilon 1}$.

multi-peaked distributions; Figure 8(g) and (h) shows the distributions for two such residues, histidine 12 in the A helix and threonine 51 in the D helix. The dynamics of these helices was examined on a PS300 graphics system using the molecular graphics program HYDRA (R. E. Hubbard, unpublished results) and it was clear that deformations of the helices as a whole were leading to the large motions of these side-chains.

Examination of the trajectory on the graphics system indicated that the A helix as a whole was twisting about the helix axis during the trajectory, leading to large fluctuations in the positions of many side-chains. Some of these side-chains

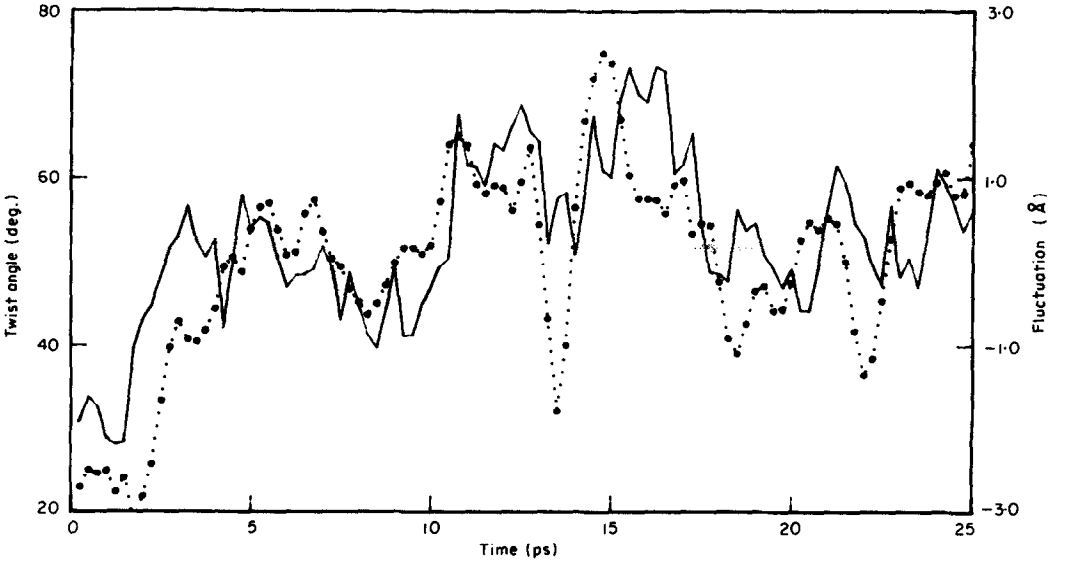

Figure 9. Correlation of the motion of the side-chain of histidine 12 with twisting of the A helix. The twist angle (as defined in the text; continuous line) and the positional fluctuation (in Å) of the $C_{\varepsilon 1}$ atom of the imidazole ring (dotted line) are plotted as a function of time. The time-series are taken from the 25 ps simulation.

undergo torsional transitions as well (e.g. Leu11, see above), leading to very complicated overall dynamics. The dihedrals of histidine 12, however, fluctuate very little during the trajectory, and most of the disorder is due to the side-chain following the twist of the helix backbone.

Using the dynamical average structure, a rigid body principal axis frame is defined for the A-helix backbone. This is the co-ordinate frame in which the moment of inertia tensor for the backbone atoms is diagonal. In this co-ordinate system, the Z-co-ordinate of an atom is its position along the helix axis and the X and Y co-ordinates specify its distance from the axis and its rotation about the helix axis. A time-series for the rotation of the C_α atom of histidine 12 about the helix axis was calculated by rotating and translating every co-ordinate set from the simulation into the helix principal axis frame defined above, and defining the twist angle as $\tan^{-1}(y/x)$. The twist about the helix axis is plotted as a function of time in Figure 9 along with the time-series for the positional fluctuation of the $C_{\varepsilon 1}$ atom of the imidazole ring. It is seen that the motion of the ring atom follows the twist of the backbone about the helix axis.

4. Conclusions

Protein refinement procedures have been tested by means of data generated by molecular dynamics simulations. It has been shown that the use of single-site isotropic models for atoms in refinement methods for proteins leads to errors in the determination of their temperature factors and average positions. These errors are smallest for atoms with low mobility, but can be very serious for atoms with temperature factors exceeding about 20 Å^2 (mean-square fluctuations exceeding about 0·75 Å^2). The magnitudes of the anharmonicity and anisotropy in myoglobin are very similar to those found for lysozyme (Ichiye & Karplus, unpublished results), suggesting that these results may be of general significance.

The neglect of solvent in the simulation is an approximation that is likely to alter the details of the dynamics, especially for the surface residues. Even though this may affect the magnitudes of the fluctuations, it should not affect the conclusion of this paper that distribution functions with multiple peaks are the most important cause of anharmonicity and anisotropy in proteins, and that this leads to temperature factors being underestimated and to the refined structure being different from the dynamical averaged structure.

The multiple peaks in the distribution function are well-separated in some residues, and alternate conformations can be picked out by difference Fourier techniques and these can be included in the refinement model. For most atoms, however, such conformations are difficult to resolve at 1·5 Å resolution. Modelling disorder is further complicated by the fact that changes in side-chain torsional angles alone might not be enough to account for the changes in the conformation of a residue.

The use of stereochemical restraints in the refinements leads to better agreement with the average dynamical structure than the use of no restraints. The inadequacy of the single-site model causes unrestrained atoms to move away from their true positions if they are close to disordered regions of the protein. The positional errors obtained on applying tight restraints on stereochemistry are not larger than those obtained using loose restraints, while the stereochemical parameters are greatly improved. Thus, refinement even at 1·5 Å resolution benefits from the use of such restraints. The restraints on temperature factor differences are too

restrictive, however, and result in a significant damping of the side-chain temperature factors.

Refinements of data generated from the simulation always converged to *R*-factors greater than 12·0% regardless of whether restraints were used or not. Higher-resolution refinements would need to be done in order to determine the most appropriate models to be used for protein refinements.

Appendix

To generate time-averaged structure factors from a molecular dynamics trajectory, the central assumption is that the simulation of the dynamics of an isolated protein molecule yields an adequate description of the probability of atomic fluctuations in a crystal. Given no information about the crystal lattice dynamics, one of two assumptions can be made about the model crystal from which the structure factors will be derived.

The first assumption, which leads to model 1, is that at every instant the crystal contains a large number of identical unit cells, each of which evolves identically in time according to the molecular dynamics trajectory. An alternative assumption is that, while the molecular dynamics simulation describes the time evolution of any one unit cell, at any instant the crystal would have unit cells which, rather than being identical, represented different configurations along the molecular dynamics trajectory; in other words, it is assumed that the unit cells are all uncorrelated.

The consequences of these assumptions can be analysed in terms of simple diffraction theory. The diffracted intensity $I(\mathbf{Q})$ at a point $\mathbf{Q}$ in reciprocal space (see eqn (1) of the main text), is proportional to the square of the Fourier transform of the electron density in the crystal:

$$I(\mathbf{Q}) = \text{const.}\ |\int \mathrm{d}\mathbf{X}\, P(\mathbf{X})\, \mathrm{e}^{i\mathbf{Q}\cdot\mathbf{X}}|^2, \qquad \text{(A1)}$$

where $P(\mathbf{X})$ is the electron density at position $\mathbf{X}$, the integral being over the whole crystal. A more convenient form is obtained by writing the integral as a sum over unit cells. Let $\rho_\alpha(\mathbf{r})$ be the electron density within unit cell α. Then, ignoring the constant term in what follows:

$$I(\mathbf{Q}) = \left| \sum_{\alpha=1}^{N} \int \mathrm{d}\mathbf{r}\, \rho_\alpha(r)\, \mathrm{e}^{i\mathbf{Q}\cdot(\mathbf{r}+\mathbf{r}_\alpha)} \right|^2, \qquad \text{(A2)}$$

where $\mathbf{X} = \mathbf{r} + \mathbf{r}_\alpha$ and the integral is now only over a single unit cell and is summed over the N unit cells in the crystal.

If there is motion in the crystal, then the density at any point is time dependent. Including the time, t, explicitly, and expanding the square, gives:

$$I(\mathbf{Q}, t) = \sum_{\alpha=1}^{N} \sum_{\alpha'=1}^{N} \mathrm{e}^{i\mathbf{Q}\cdot(\mathbf{r}_\alpha - \mathbf{r}_{\alpha'})} \times \left[\int \mathrm{d}\mathbf{r} \int \mathrm{d}\mathbf{r}'\, \rho_\alpha(\mathbf{r}, t)\rho_{\alpha'}(\mathbf{r}', t)\, \mathrm{e}^{i\mathbf{Q}\cdot(\mathbf{r}-\mathbf{r}')} \right]. \qquad \text{(A3)}$$

The time-averaged intensity, which is the quantity of interest, is:

$$\langle I(\mathbf{Q})\rangle = \sum_{\alpha=1}^{N} \sum_{\alpha'=1}^{N} \mathrm{e}^{i\mathbf{Q}\cdot(\mathbf{r}_\alpha - \mathbf{r}_{\alpha'})} \times \int \mathrm{d}\mathbf{r} \int \mathrm{d}\mathbf{r}' \langle \rho_\alpha(\mathbf{r}, t)\rho_{\alpha'}(\mathbf{r}', t)\rangle\, \mathrm{e}^{i\mathbf{Q}\cdot(\mathbf{r}_\alpha - \mathbf{r}_{\alpha'})} \qquad \text{(A4)}$$

The two models will now be considered separately.

(a) *Model 1*

Since the unit cells are identical at every instant:

$$\rho_\alpha(\mathbf{r}, t) = \rho_{\alpha'}(\mathbf{r}, t), \quad \text{for all } \alpha, \alpha' \text{ and } t. \qquad \text{(A5)}$$

Therefore,

$$\langle I(\mathbf{Q})\rangle = \sum_{\alpha=1}^{N} \sum_{\alpha'=1}^{N} \mathrm{e}^{i\mathbf{Q}\cdot(\mathbf{r}_\alpha - \mathbf{r}')} \times \int \mathrm{d}\mathbf{r} \int \mathrm{d}\mathbf{r}' \langle \rho(\mathbf{r}, t)\rho(\mathbf{r}', t)\rangle\, \mathrm{e}^{i\mathbf{Q}\cdot(\mathbf{r}-\mathbf{r}')} \qquad \text{(A6)}$$

or:

$$\Rightarrow \langle I(\mathbf{Q})\rangle = \left\langle \left| \sum_{\alpha=1}^{N} \mathrm{e}^{i\mathbf{Q}\cdot\mathbf{r}_\alpha} \int \mathrm{d}\mathbf{r}\, \rho(\mathbf{r})\, \mathrm{e}^{i\mathbf{Q}\cdot\mathbf{r}} \right|^2 \right\rangle \qquad \text{(A7)}$$

and:

$$\Rightarrow \langle I(\mathbf{Q})\rangle = 2\pi^3 N \sum_{\mathbf{H}} \delta(\mathbf{Q} - 2\pi\mathbf{H}) \langle |F(\mathbf{Q}, t)|^2\rangle, \qquad \text{(A8)}$$

where $F(\mathbf{Q}, t)$ is the structure factor, i.e. the Fourier transform of the contents of one of the N identical unit cells, at time t, and $\mathbf{H}$ is a reciprocal lattice point. The steps in going from equation (A7) to (A8) are explained in section (b), below. Thus model 1 predicts non-zero intensity only at the reciprocal lattice points, $\mathbf{H}$, and:

$$\langle I(\mathbf{Q})\rangle \propto \langle |F(\mathbf{Q}, t)|^2\rangle : \text{at } \mathbf{Q} = 2\pi\mathbf{H}. \qquad \text{(A9)}$$

This quantity can be evaluated readily from the molecular dynamics trajectory by Fourier transforming the electron density at each time-step and averaging the squares of the calculated structure factors.

(b) *Model 2*

In this case, no two unit cells are identical, but the fact that different unit cells are uncorrelated can be used to simplify the expression. The instantaneous density can be written as:

$$\rho_\alpha(\mathbf{r}, t) = \langle \rho_\alpha(\mathbf{r})\rangle + \Delta\rho_\alpha(\mathbf{r}, t), \qquad \text{(A10)}$$

where $\langle \rho_\alpha(\mathbf{r})\rangle$ is the time-averaged electron density at $\mathbf{r}$ and $\Delta\rho(\mathbf{r}, t)$ is the instantaneous fluctuation in the density. The time-averaged density is constant from unit cell to unit cell assuming a homogeneous crystal, and the α dependence in the first term can be dropped. Then:

$$I(\mathbf{Q}) = \sum_{\alpha=1}^{N} \sum_{\alpha'=1}^{N} \mathrm{e}^{i\mathbf{Q}\cdot(\mathbf{r}_\alpha - \mathbf{r}_{\alpha'})} \times \int \mathrm{d}\mathbf{r} \int \mathrm{d}\mathbf{r}' \langle \rho(\mathbf{r})\rangle\langle \rho(\mathbf{r}')\rangle\, \mathrm{e}^{i\mathbf{Q}\cdot(\mathbf{r}-\mathbf{r}')} + \sum_{\alpha=1}^{N} \sum_{\alpha'=1}^{N} \mathrm{e}^{i\mathbf{Q}\cdot(\mathbf{r}_\alpha - \mathbf{r}_{\alpha'})} \times \int \mathrm{d}\mathbf{r} \int \mathrm{d}\mathbf{r}' \langle \Delta\rho_\alpha(\mathbf{r}, t)\Delta\rho_{\alpha'}(\mathbf{r}', t)\rangle\, \mathrm{e}^{i\mathbf{Q}\cdot(\mathbf{r}-\mathbf{r}')}. \qquad \text{(A11)}$$

Naming the first term I_1 and the second I_2, it can now be shown that for a homogeneous crystal, I_1 is $0(N^2)$, while I_2 is $0(N)$.

We have:

$$I_1 = \sum_{\alpha=1}^{N} \sum_{\alpha'=1}^{N} e^{i\mathbf{Q}\cdot(\mathbf{r}_\alpha - \mathbf{r}_{\alpha'})} \times \int d\mathbf{r} \int d\mathbf{r}' \langle \rho(\mathbf{r}) \rangle \langle \rho(\mathbf{r}') \rangle e^{i\mathbf{Q}\cdot(\mathbf{r}-\mathbf{r}')} \quad \text{(A12)}$$

or:

$$I_1 = \sum_{\alpha=1}^{N} \sum_{\alpha'=1}^{N} e^{i\mathbf{Q}\cdot(\mathbf{r}_\alpha - \mathbf{r}_{\alpha'})} |\langle F(\mathbf{Q}) \rangle|^2, \quad \text{(A13)}$$

where $\langle F(\mathbf{Q}) \rangle$ is the Fourier transform of the time-averaged electron density in a unit cell. If **a**, **b** and **c** are the direct unit cell vectors, and **a***, **b*** and **c*** are the reciprocal unit cell vectors, then

$$\mathbf{Q} = 2\pi(h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*)$$

(h, k and l are not, in general, integral; see eqn (1) of the main text) and

$$\mathbf{r}_\alpha = \beta\mathbf{a} + \gamma\mathbf{b} + \delta\mathbf{c},$$

where β, γ and δ are integers. Then equation (A13) can be written as:

$$I_1(\mathbf{Q}) = |\langle F(\mathbf{Q}) \rangle|^2 \left| \sum_{\beta=1}^{N_x} e^{2\pi i h\beta} \right|^2 \left| \sum_{\gamma=1}^{N_y} e^{2\pi i k\gamma} \right|^2 \left| \sum_{\delta=1}^{N_z} e^{2\pi i l\delta} \right|^2, \quad \text{(A14)}$$

where N_x, N_y and N_z are the numbers of unit cells along the three cell edges, and $N_x \times N_y \times N_z = N$. By expanding the exponentials and simplifying, this leads to:

$$I_1(\mathbf{Q}) = \frac{\sin^2(\pi h N_x)}{\sin^2(\pi h)} \frac{\sin^2(\pi k N_y)}{\sin^2(\pi k)} \times \frac{\sin^2(\pi l N_z)}{\sin^2(\pi l)} |\langle F(\mathbf{Q}) \rangle|^2. \quad \text{(A15)}$$

The function:

$$\frac{\sin^2(\pi h N_x)}{\sin^2(\pi h)} \frac{\sin^2(\pi k N_y)}{\sin^2(\pi k)} \frac{\sin^2(\pi l N_z)}{\sin^2(\pi l)}$$

is periodic in h, k, l and is called the Laue interference function. It is very sharply peaked around:

$$h = \text{integer}, \; k = \text{integer}, \; l = \text{integer}. \quad \text{(A16)}$$

These are the Bragg conditions that define the reciprocal lattice points (h, k, l), and so the first conclusion is that I_1 corresponds to Bragg scattering in that it is proportional to the square of the Fourier transform of the time-averaged electron density, and in that it is sharply peaked at the reciprocal lattice points. The Laue interference function must be evaluated in order to compare the relative magnitudes of I_1 and I_2. If the Laue interference function is evaluated exactly at a reciprocal lattice point, its value is N^2. This is because:

$$\lim_{h \to \text{integer}} \frac{\sin^2(\pi h N_x)}{\sin^2(\pi h)} = N_x^2. \quad \text{(A17)}$$

It is more common to write I_1 in terms of its behaviour under integration. This may be done by looking at its behaviour near one diffraction spot ($h = 0$, $k = 0$, $l = 0$) and recognizing that the Laue interference function is periodic. If we integrate I_1 over a volume small compared to the reciprocal lattice cell volume (i.e. integrate near a lattice point, presumably as is done by a diffractometer) and treat $|F(\mathbf{Q})|^2$ as constant, we get:

$$\int_{\mathbf{Q}-\Delta\mathbf{Q}}^{\mathbf{Q}+\Delta\mathbf{Q}} d\mathbf{Q}\, I_1(\mathbf{Q}) = |\langle F(\mathbf{Q}) \rangle|^2 \times \int_{-\Delta h}^{+\Delta h} \frac{\sin^2(\pi h N_x)}{\sin^2(\pi h)} 2\pi dh \times \int_{-\Delta k}^{+\Delta k} \frac{\sin^2(\pi k N_y)}{\sin^2(\pi k)} 2\pi dk \times \int_{-\Delta l}^{+\Delta l} \frac{\sin^2(\pi l N_z)}{\sin^2(\pi l)} 2\pi dl. \quad \text{(A18)}$$

For large values of N, each of the $\sin^2$ terms rapidly goes to zero as h, k and l deviate from zero, i.e. the terms go to zero for:

$$|h| > \frac{1}{N_x}, \quad |k| > \frac{1}{N_y} \quad \text{and} \quad |l| > \frac{1}{N_z}.$$

So, in the limit of large N_x:

$$\int_{-\Delta h}^{+\Delta h} \frac{\sin^2(\pi h N_x)}{\sin^2(\pi h)} 2\pi dh = \int_{-\infty}^{+\infty} \frac{\sin^2(\pi h N_x)}{(\pi h)^2} 2\pi dh = 2\pi N_x \quad \text{(A19)}$$

and similarly for the integrals over k and l. Thus, equation (A18) reduces to:

$$\int_{\mathbf{Q}-\Delta\mathbf{Q}}^{\mathbf{Q}+\Delta\mathbf{Q}} d\mathbf{Q}\, I_1(\mathbf{Q}) = (2\pi)^3 N |\langle F(2\pi\mathbf{H}) \rangle|^2. \quad \text{(A20)}$$

Equation (A20) implies that $I_1(\mathbf{Q})$ can be written as:

$$I_1(\mathbf{Q} = (2\pi)^3 N \sum_{\mathbf{H}} \delta(\mathbf{Q} - 2\pi\mathbf{H}) |\langle F(\mathbf{Q}) \rangle|^2, \quad \text{(A21)}$$

where $\delta(\mathbf{Q} - 2\pi\mathbf{H})$ is the Dirac delta function and is non-zero only at the reciprocal lattice points.

To evaluate I_2, the fact that the dynamics of atoms in different unit cells are uncorrelated is used to simplify the expression in the following way:

$$\begin{aligned} I_2(\mathbf{Q}) &= \sum_{\alpha=1}^{N} \sum_{\alpha'=1}^{N} e^{i\mathbf{Q}\cdot(\mathbf{r}_\alpha - \mathbf{r}_{\alpha'})} \int d\mathbf{r} \int d\mathbf{r}' \\ &\quad \times \langle \Delta\rho_\alpha(\mathbf{r}, t) \Delta(\rho(\mathbf{r}', t) \rangle e^{i\mathbf{Q}\cdot(\mathbf{r}-\mathbf{r}')} \\ &= \sum_{\alpha=1}^{N} \int d\mathbf{r} \int d\mathbf{r}' \langle \Delta\rho_\alpha(\mathbf{r}, t) \Delta\rho_\alpha(\mathbf{r}', t) \rangle e^{i\mathbf{Q}\cdot(\mathbf{r}-\mathbf{r}')} \\ &\quad + \sum_{\alpha=1}^{N} \sum_{\alpha' \neq \alpha}^{N} e^{i\mathbf{Q}\cdot(\mathbf{r}_\alpha - \mathbf{r}_{\alpha'})} \int d\mathbf{r} \int d\mathbf{r}' \\ &\quad \times \langle \Delta\rho_\alpha(\mathbf{r}, t) \Delta\rho_{\alpha'}(\mathbf{r}', t) \rangle e^{i\mathbf{Q}\cdot(\mathbf{r}-\mathbf{r}')}. \end{aligned} \quad \text{(A22)}$$

When different unit cells are uncorrelated, the term $\langle \Delta\rho_\alpha(\mathbf{r}, t) \Delta\rho_{\alpha'}(\mathbf{r}', r) \rangle$ is zero for $\alpha \neq \alpha'$. Hence:

$$I_2(\mathbf{Q}) = N \langle |\Delta F(\mathbf{Q})|^2 \rangle, \quad \text{(A23)}$$

where

$$\Delta F(\mathbf{Q}) = \int d\mathbf{r}' \Delta\rho(\mathbf{r})\, e^{i\mathbf{Q}\cdot\mathbf{r}}.$$

So now we have:

$$I(\mathbf{Q}) = (2\pi)^3 N \sum_{\mathbf{H}} \delta(\mathbf{Q}-2\pi\mathbf{H})|\langle F(\mathbf{Q})\rangle|^2 + N\langle|\Delta F(\mathbf{Q})|^2\rangle. \quad \text{(A24)}$$

The presence of the delta function ensures that the first term dominates the scattering at the reciprocal lattice points.

To summarize, there are two simple ways in which structure factors calculated from a simulation of a single molecule can be used to derive time-averaged diffraction intensities for crystals. In one case, the structure factors are first squared and then averaged, and in the other the structure factors are first averaged and then squared. The problem with the first method is that it cannot be related directly to the average electron density, and so attempts to use conventional crystallographic methods to arrive at a model for the intensities calculated that way may not be very successful. When averaged intensities are calculated using both methods and compared, a crystallographic *R*-factor (eqn (9): main text) of 36% is obtained. A refinement of a model structure against the $\langle|F|\rangle^2$ data (i.e. assuming perfectly correlated unit cells) using the average co-ordinates from the dynamics as the starting model was made. Three positional parameters and a temperature factor were refined for each atom, using data between 10·0 Å and 2·0 Å. The final *R*-factor did not drop below 24%, which is actually worse than the final *R*-factor for many proteins at 2·0 Å resolution (for well-refined structures these are around 15% to 17%). The temperature factors obtained from this refinement were about a factor of 10 lower than those calculated directly from the simulation.

These results provide evidence for the fact that of the two limiting models, that assuming no correlation of unit cells is much closer to the truth than one assuming perfect correlation of the molecular motion in different unit cells. However, some correlation probably does exist, and it would be of considerable interest to examine this question in more detail.

We thank Toshiko Ichiye for many stimulating discussions about protein refinement and Rod Hubbard for help with the display program HYDRA. This work has been supported by grants from the National Institutes of Health NIHGM26788 (to G.A.P.) and the National Science Foundation (to M.K.). R.M.L. is an Alfred P. Sloan Fellow and a recipient of an N.I.H. Research Career Development award.

References

Agarwal, R. C. (1978). *Acta Crystallogr. sect. A*, **34**, 791–809.

Amoros, J. L. & Amoros, M. (1968). *Molecular Crystals: Their Transforms and Diffuse Scattering*, John Wiley and Sons, New York.

Artymiuk, P. J., Blake, C. C. F., Grace, D. E. P., Oatley, S. T., Phillips, D. C. & Sternberg, M. J. E. (1979). *Nature (London)*, **280**, 563–568.

Blundell, T. & Johnson, L. N. (1976). *Protein Crystallography*, Academic Press, New York.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). *J. Comp. Chem.* **4**, 187–217.

Diamond, R. (1971). *Acta Crystallogr. sect. A*, **27**, 436–452.

Dunitz, J. (1979). *X-ray Analysis and the Structure of Organic Molecules*, Cornell University Press, Ithaca, N.Y.

Frauenfelder, H., Petsko, G. A. & Tsernoglou, D. (1979). *Nature (London)*, **280**, 558–563.

Hartmann, H., Parak, F., Steigmann, W., Petsko, G. A., Ponzi, D. R. & Frauenfelder, H. (1982). *Proc. Nat. Acad. Sci., U.S.A.* **79**, 4967–4971.

Hendrickson, W. A. (1980). In *Refinement of Protein Structures; Proceedings of the Daresbury Study Weekend* (Machin, P. A. & Elder, M., eds), pp. 1–8, Science and Engineering Research Council, Daresbury Laboratory, U.K.

Hendrickson, W. A. (1985). *Meth. Enzymol.* **115**, 252–270.

Hendrickson, W. A. & Konnert, J. (1980). In *Computing in Crystallography* (Diamond, R., Ramasheshan, S. & Venkatesan, K., eds), pp. 13.01–13.23, Indian Institute of Science, Bangalore.

Hirshfeld, F. L. & Rabinovich, D. (1973). *Acta Crystallogr. sect. A*, **29**, 510–513.

International Tables for X-Ray Crystallography (1974). Ibers, J. & Hamilton, W. C., eds, vol. 4, International Union of Crystallography, The Kynoch Press, Birmingham.

James, R. W. (1948). *The Optical Principles of X-Ray Diffraction* (Reissued 1982), Ox Bow Press, Woodbridge, Ct.

Jones, T. A. (1982). In *Computational Crystallography* (Sayre, D., ed.), pp. 3303–3317, Clarendon, Oxford.

Karplus, M. (1981). *Ann. N.Y. Acad. Sci.* **367**, 407–418.

Karplus, M. & McCammon, J. A. (1981). *C.R.C. Crit. Rev. Biochem.* **9**, 293–349.

Karplus, M. & McCammon, J. A. (1983). *Annu. Rev. Biochem.* **53**, 263–300.

Konnert, J. H. (1976). *Acta Crystallogr. sect. A*, **32**, 614–617.

Konnert, J. H. & Hendrickson, W. A. (1980). *Acta Crystallogr. sect. A*, **36**, 344–349.

Levy, R. M. & Keepers, J. (1986). *Comments on Molecular and Cellular Biophysics*, in the press.

Levy, R. M., Karplus, M. & Wolynes, P. (1981). *J. Amer. Chem. Soc.* **103**, 5998.

Levy, R. M., Sheridan, R. P., Keepers, J., Dubey, G. S. Swaminathan, S. & Karplus, M. (1985). *Biophys. J.* **48**, 509–518.

Mao, B., Pear, M. R. & McCammon, J. A. (1982). *Biopolymers*, **21**, 1979–1989.

McCammon, J. A., Gelin, B. R. & Karplus, M. (1977). *Nature (London)*, **267**, 585–590.

Moore, F. H. (1963). *Acta Crystallogr.* **16**, 1169–1175.

Morokuma, K. & Karplus, M. (1971). *J. Chem. Phys.* **55**, 63–75.

Northrup, S. H., Pear, M. R., Morgan, J. D., McCammon, J. A. & Karplus, M. (1981). *J. Mol. Biol.* **153**, 1087–1090.

Petsko, G. A. & Ringe, D. (1984). *Annu. Rev. Biophys. Bioeng.* **13**, 331–371.

Phillips, S. E. V. (1980). *J. Mol. Biol.* **142**, 531–554.

Sheriff, S., Hendrickson, W. A., Stenkamp, R. E., Sieker, L. C. & Jensen, L. H. (1985). *Proc. Nat. Acad. Sci., U.S.A.* **82**, 1104–1107.

Stewart, R. F. & Feil, D. (1980). *Acta Crystallogr. sect. A*, **36**, 503–509.

Takano, T. & Dickerson, R. E. (1981). In *Interaction between Iron and Protein in Oxygen and Electron Transport* (Ho, C., ed.), Elsevier/North-Holland, New York.

Ten Eyck, L. F. (1973). *Acta Crystallogr. sect. A*, **29**, 183–191.

Ten Eyck, L. F. (1977). *Acta Crystallogr. sect. A*, **33**, 486–492.

Teeter, M. M. & Hendrickson, W. A. (1979). *J. Mol. Biol.* **127**, 219–233.

van Gunsteren, W. F. & Karplus, M. (1982*a*). *Biochemistry*, **21**, 2259–2274.

van Gunsteren, W. F. & Karplus, M. (1982*b*). *Macromolecules*, **15**, 1528–1544.

van Gunsteren, W. F., Berendsen, H. J. C., Hermans, J., Hol, W. G. J. & Postma, J. P. M. (1983). *Proc. Nat. Acad. Sci., U.S.A.* **80**, 4315–4319.

Watenpaugh, K. D., Sieker, L. C. & Jensen, L. H. (1980). *J. Mol. Biol.* **138**, 615–633.

Willis, B. T. M. & Pryor, W. (1975). *Thermal Vibrations in Crystallography*, Cambridge University Press. Cambridge.

Wilson, K. S., Stura, E. A., Wild, D. L., Todd, R. J., Stuart, D. I., Babu, Y. S., Jenkins, J. A., Standing, T. S., Johnson, L. N., Fourme, R., Kann, R., Gadet, A., Bartels, K. S. & Bartunik, H. D. (1983). *J. Appl. Crystallogr.* **16**, 28–41.

Wlodawer, A., Walter, J., Huber, R. & Sjolin, L. (1984). *J. Mol. Biol.* **180**, 301–329.

Woolfson, M. M. (1970). *An Introduction to X-Ray Crystallography*, Cambridge University Press, Cambridge.

Xuong, N. H., Freer, S. T., Hamlin, R., Nielsen, C. & Vernon, W. (1978). *Acta Crystallogr. sect. A*, **34**, 289–296.

Yu, H., Karplus, M. & Hendrickson, W. (1985). *Acta Crystallogr. sect. B*, **41**, 191–201.

Zucker, U. H. & Schulz, H. (1982). *Acta Crystallogr. sect. A*, **38**, 563–568.

Edited by A. Klug