January 2007

Anisotropic atomic motions in high-resolution protein crystallography molecular dynamics simulations

Conrad J. Burden
ANU

Aaron J. Oakley
University of Wollongong, aarono@uow.edu.au

Anisotropic atomic motions in high-resolution protein crystallography molecular dynamics simulations

Abstract

Molecular dynamics (MD) simulations using empirical force fields are popular for the study of proteins. In this work, we compare anisotropic atomic fluctuations in nanosecond-timescale MD simulations with those observed in an ultra-high-resolution crystal structure of crambin. In order to make our comparisons, we have developed a compact graphical technique for assessing agreement between spatial atomic distributions determined by MD simulations and observed anisotropic temperature factors.

Keywords

anisotropic, dynamics, atomic, simulations, motions, high, resolution, protein, crystallography, molecular, CMMB

Publication Details

Burden, C. J. & Oakley, A. J. (2007). Anisotropic atomic motions in high-resolution protein crystallography molecular dynamics simulations. Physical Biology, 4 (2), 79-90.

Phys. Biol. **4** (2007) 79–90 doi:10.1088/1478-3975/4/2/002

Anisotropic atomic motions in high-resolution protein crystallography molecular dynamics simulations

Conrad J Burden[1,2] **and Aaron J Oakley**[3]

[1] Centre for Bioinformation Science, Mathematical Sciences Institute, Australian National University, Canberra, ACT 0200, Australia
[2] John Curtin School of Medical Research, Australian National University, Canberra, ACT 0200, Australia
[3] Research School of Chemistry, Australian National University, Canberra, ACT 0200, Australia

E-mail: Conrad.Burden@anu.edu.au

Received 5 February 2007
Accepted for publication 21 May 2007
Published 11 June 2007
Online at stacks.iop.org/PhysBio/4/79

Abstract
Molecular dynamics (MD) simulations using empirical force fields are popular for the study of proteins. In this work, we compare anisotropic atomic fluctuations in nanosecond-timescale MD simulations with those observed in an ultra-high-resolution crystal structure of crambin. In order to make our comparisons, we have developed a compact graphical technique for assessing agreement between spatial atomic distributions determined by MD simulations and observed anisotropic temperature factors.

Abbreviations

ATF	anisotropic temperature factor
CHARMM	chemistry at Harvard macromolecular mechanics
ESD	estimated standard deviation
MD	molecular dynamics
NAMD	nanoscale molecular dynamics
PDB	Protein Data Bank
RMS	root mean square
TIP3P	transferable intermolecular potential three-point model
UVS	unit variance spheroid

1. Introduction

Molecular dynamics simulations are an increasingly popular tool for investigating the energetics and mechanics of biomolecules. Protein molecules fluctuate due to thermal motion, and some undergo structural rearrangements as part of their normal function. Thanks to experimental techniques such as atomic force microscopy and theoretical approaches such as molecular dynamics (MD) simulations, there is an increasing awareness of the dynamic nature of proteins.

The Protein Data Bank (PDB) [1] is a database of experimentally determined three-dimensional structures of biological macromolecules. The PDB file of a protein structure determined by crystallography typically contains four parameters for every atom: measured mean coordinates $\mathbf{r}_0 = (x_0, y_0, z_0)^{\mathrm{T}}$ of atoms and the Debye–Waller (or B) factor, related to the motions of the atom:

$$B = \frac{8\pi^2}{3}\langle(\mathbf{r} - \mathbf{r}_0)^2\rangle, \tag{1}$$

where $\langle(\mathbf{r} - \mathbf{r}_0)^2\rangle$ is the linear mean square displacement of the atom about its mean postion. This is an 'isotropic' B-factor insofar as it assumes that the atomic fluctuations are uniform in all directions. In the isotropic B-factor model, the crystallographic structure factor F_{calc} is of the form

$$F_{\mathrm{calc}}(\mathbf{h}) = \sum_j f_j \exp\left(-\frac{1}{4}B_j\mathbf{h}^{\mathrm{T}}\mathbf{h}\right)\exp(2\pi i\mathbf{h}^{\mathrm{T}}\mathbf{r}_{0j}), \tag{2}$$

where three mean coordinates $\mathbf{r}_{0j}$ and one isotropic B-factor B_j are defined for each atom j.

Atomic displacements result from thermal vibration (dynamic disorder), different discrete conformations of molecules in different unit cells (static disorder), lattice defects and lattice vibrations (phonons). Thanks to synchrotron

radiation, an increasing number of biomolecular structures are being determined at atomic resolution. When the x-ray data extends beyond Bragg spacings of about 1.2 Å, it is possible to give a more sophisticated description of the thermal motions of atoms. The extra information resulting from anisotropic displacements of atoms is recorded in PDB files as a set of six numbers for each atom known as anisotropic temperature factors (ATFs) $U_{11}, U_{22}, U_{33}, U_{12}, U_{13}$ and U_{23}. These numbers are elements of the symmetric variance–covariance matrix of the atom's spatial probability distribution.

To a first approximation, and for those atoms which only exist in a single conformer, the probability distribution for finding an atom at the location $\mathbf{r} = (x, y, z)^{\mathrm{T}}$ is assumed to be given by a trivariate normal distribution with density function

$$\phi(\mathbf{r}) = \frac{1}{(2\pi)^{3/2}(\det U)^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{r}-\mathbf{r}_0)^{\mathrm{T}} U^{-1}(\mathbf{r}-\mathbf{r}_0)\right], \tag{3}$$

where the covariance matrix U is a 3×3 real symmetric matrix whose elements U_{ij} are the ATFs[4]. From equation (1), it can be shown that the isotropic temperature factor B is related to the elements of U by

$$B = \frac{8\pi^2}{3}(U_{11} + U_{22} + U_{33}), \tag{4}$$

where the angle brackets in equation (1) denote the expectation value with respect to the distribution equation (3).

The matrix U can be decomposed uniquely into the product

$$U = R \begin{pmatrix} {\sigma_1}^2 & & \\ & {\sigma_2}^2 & \\ & & {\sigma_3}^2 \end{pmatrix} R^{\mathrm{T}}, \tag{5}$$

where ${\sigma_1}^2 \geqslant {\sigma_2}^2 \geqslant {\sigma_3}^2 \geqslant 0$ are the (real) eigenvalues of U and $R \in SO(3)$ is a real orthogonal unimodular 3×3 matrix, that is, a rotation matrix in three dimensions satisfying $R^{\mathrm{T}}R = RR^{\mathrm{T}} = I$ and $\det R = 1$. The distribution $\phi(\mathbf{r})$ can be visualized as a set of concentric spheroids of constant variance. The spheroid at unit variance has principal axes of lengths $2\sigma_1, 2\sigma_2$ and $2\sigma_3$, whose orientation is obtained by applying the rotation matrix R to unit vectors $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ lying along the x, y and z axes. We shall refer to this spheroid as the unit variance spheroid (UVS). We define $\mathbf{a}_i = R\mathbf{e}_i$, $i = 1, 2, 3$, to be the unit vectors lying along the principal axes of the UVS (see figure 1).

Atom trajectories within proteins can be determined independently from a theoretical perspective from MD simulations. From these trajectories predictions of mean atomic coordinates and isotropic and anisotropic temperature factors can be inferred. The purpose of this paper is to propose a quantitative measure and technique for visualizing graphically the goodness of fit between anisotropic temperature factors determined from experiment and those from MD simulations. As far as we are aware, the only previous analysis of this sort is that of Komeiji *et al* [2] who analyse the human lysozyme protein. We apply our analysis to the ultra-high-resolution data for the crambin protein [3] (see figure 2).

[4] The PDB files use a convention of listing the ATFs multiplied by 10^4.

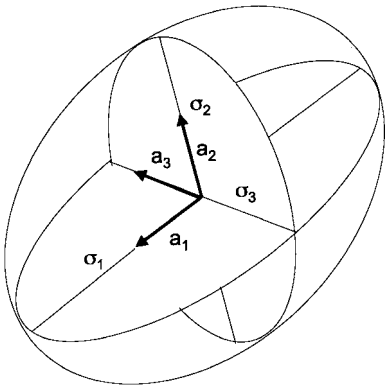

Figure 1. The unit variance spheroid (UVS) of a trinomial probability density distribution.

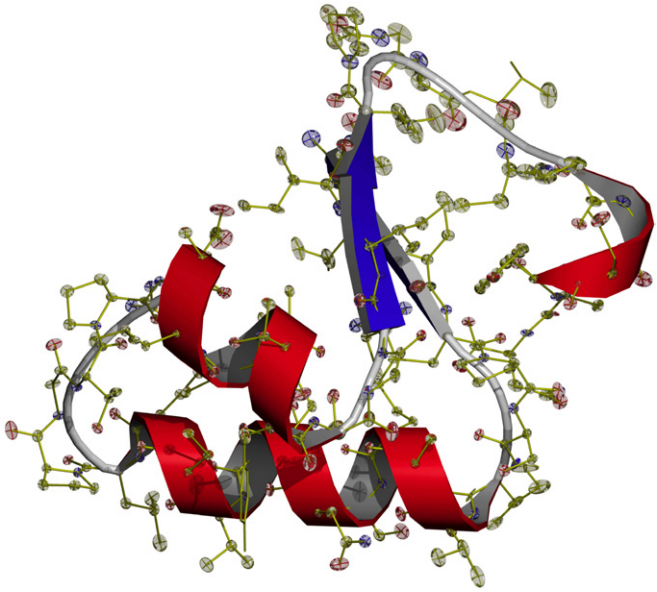

Figure 2. Structure of the crambin molecule determined from high-resolution x-ray diffraction data [3].

2. Temperature factors from MD trajectories

We next describe our MD simulation trajectories for the crambin molecule. For all simulations, NAMD 2.5 [4, 5] was used. PDB model 1GVY was used for the starting coordinates of the protein moiety in all simulations. The CHARMM22 force field [6] was employed. Both molecules in the asymmetric unit were simulated, with periodic boundary conditions employed to recreate the crystal environment. Non-bonded interactions were scaled linearly to zero from 10 to 12 Å. The particle mesh Ewald method was used for long-range electrostatic calculations. Energy minimization (10 000 steps) was performed prior to Langevin dynamics at 100 K. The integration time-step was 1 fs and simulations of 10 ns were run in all cases. The last 8 ns were used for ATF calculations. Frames in each trajectory were RMS fitted to their starting coordinates in order to remove rigid body rotational and translational components. We are confident that

the MD simulation has reached convergence, as increasing the time period used for calculations from less than 3 ns to 8 ns showed no significant change in the results presented below.

While first, second and sometimes higher-shell water molecules are observed at discrete locations on the surface of proteins, the 'bulk solvent' regions have flat, continuous electron density which is not modelled explicitly. This is due to the presence of networks of solvent molecules that vary between unit cells, the observed electron density being the average over all unit cells. We have modelled the solvent regions using explicit TIP3P water molecules and locally enhanced sampling to simulate different networks of water molecules. Fifteen copies of each water molecule are included. There are no interactions between different copies of a given water molecule and interactions with the protein are scaled by a factor of $1/15$.

Mean coordinates and the variance–covariance matrices were estimated from trajectories $\mathbf{r}(t)$, $t_1 = 2000$ ps $\leqslant t \leqslant t_2 = 10\,000$ ps at time intervals of 10 ps using the usual unbiased estimators for means, variances and covariances, namely

$$\bar{\mathbf{r}}^{\mathrm{MD}} = \frac{1}{N}\sum_{t=t_1}^{t_2}\mathbf{r}(t), \tag{6}$$

$$U_{ij}^{\mathrm{MD}} = \frac{1}{N-1}\sum_{t=t_1}^{t_2}\left(r_i(t) - \bar{r}_i^{\mathrm{MD}}\right)\left(r_j(t) - \bar{r}_j^{\mathrm{MD}}\right), \quad i, j = 1, 2, 3, \tag{7}$$

where $N = t_2 - t_1 + 1$ is the number of time points in the trajectory.

Ichiye and Karpus [7] have analysed positional probability density functions for atomic fluctuations determined from a molecular dynamics simulation of hen egg-white lysozyme. By estimating the skewness and kurtosis from sample trajectory points, they concluded that the majority of atom trajectories are well approximated by a trivariate normal distribution. The greatest deviations from a trivariate normal distribution tend to occur for sidechain atoms, and are caused by probability density functions with multiple peaks in the direction of the local principal axis. This suggests atomic trajectories which make stochastic transitions between different local conformations.

We have used the Shapiro–Wilk normality test [8] to analyse samples taken at 10 ps intervals of our simulated crambin atom coordinates projected onto the principal axes of the variance–covariance matrix for each atom. P-values obtained from this test will be uniformly distributed on the interval $[0, 1]$ if independently and identically distributed data are sampled from a Gaussian distribution, whereas data independently sampled from a non-Gaussian distribution will produce P-values skewed heavily towards zero. Figure 3 plots the histograms of Shapiro–Wilk P-values for coordinates of all 327 atoms in the crambin molecule. Coordinates projected onto the smallest principal axis of each atom appear, in general, to be well approximated by Gaussian distributions. The histogram for the longest principal axis, however, shows a spike in the lowest one-percentile, indicating a subset of atoms whose trajectories are non-Gaussian along this axis. Figure 4 plots those MD simulation trajectories with P-values less than 0.0015. Certain atoms exhibit clear signs of bistability, notably the oxygen in residue 9 (alanine), atoms in residue 30 (threonine) and the nitrogen in residue 32 (cysteine). A smaller number of atoms show signs of bistability along the secondary axis.

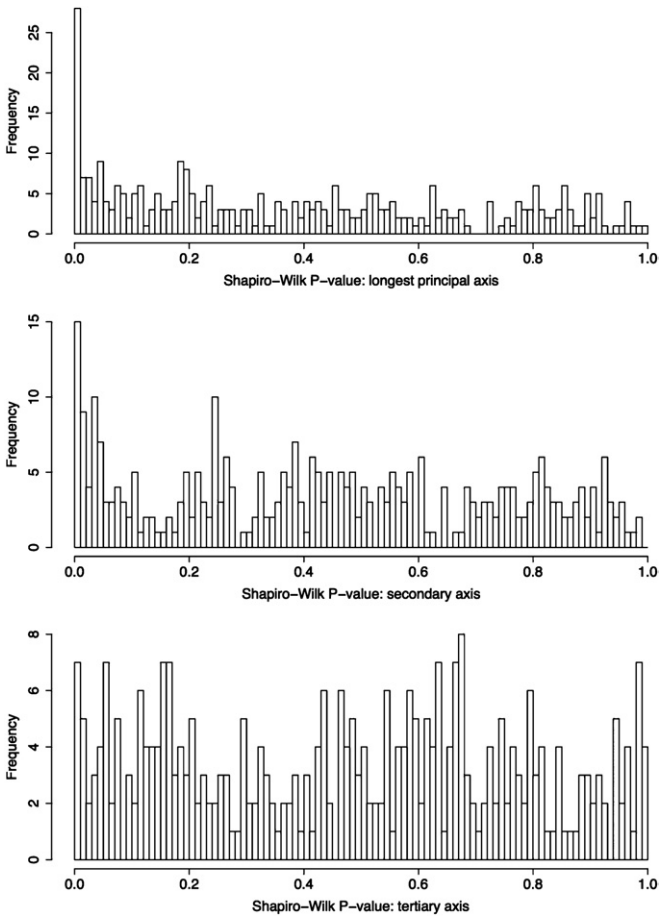

Figure 3. Histograms of P-values obtained from the Shapiro–Wilk test applied to trajectory coordinates from the crambin MD simulation. For each atom, local coordinates are rotated to the principal axes of the trajectory's variance–covariance matrix. The top histogram refers to the longest principal axis, the middle histogram to the second longest and the third histogram to the shortest principal axis.

Figure 4 also shows a single deviant point in the trajectory of the carbon-γ_1 atom in residue 34 (isoleucine). Close inspection of the trajectory and 3D rendering of this residue showed a transient rotation of the carbon-γ_1 and carbon-δ atoms at this point in the trajectory. The carbon-δ does not show up in figure 4 as the time spent in the rotated configuration was not long enough to register the direction of the transient rotation to be the largest principal axis of the anisotropic temperature factor matrix for this atom.

In the following analysis, we shall refer to the quantities U_{ij}^{MD} defined by equation (7) and quantities B^{MD} derived via equation (4) as anisotropic and isotropic temperature factors, respectively, from our MD simulation. For the majority

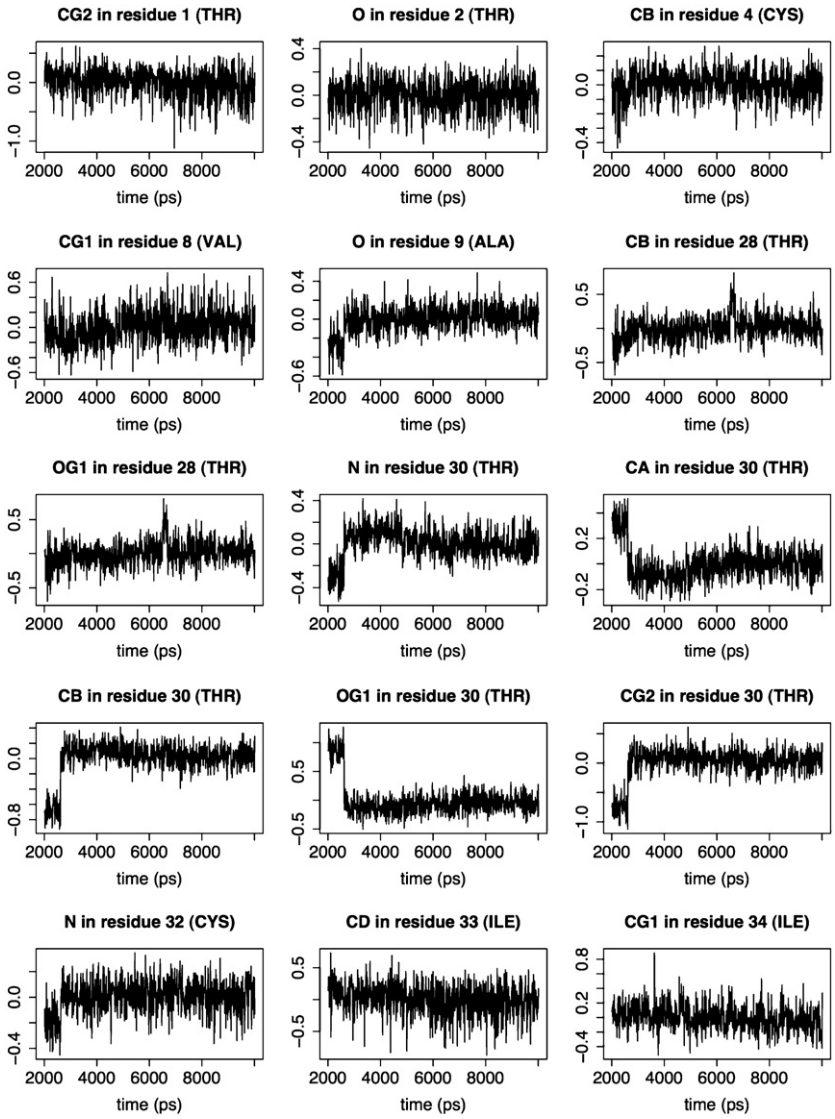

Figure 4. Deviations (in Å) from the mean position of MD simulation trajectories for those atoms contributing to the lowest 0.15-percentile of P-values in the upper histogram of figure 3. In each case the trajectory is projected onto the longest principal axis of the trajectory's variance–covariance matrix.

of atoms in the crambin MD simulation the trajectories do, to a very good approximation, generate a trivariate normal distribution, and the UVS of figure 1 is an accurate description of the shape of the probability density function. For the subset of atoms represented in figure 4 corresponding to bistable trajectories, however, the usefulness of figure 1 is restricted to indicating the orientation and lengths of the principal axes of the variance–covariance matrix.

3. Isotropic temperature factors

Figure 5 plots iostropic temperature factors B for each atom in the crambin molecule as measured by x-ray crystallography and as determined from the MD simulation. For most atoms, the MD simulation gives isotropic temperature factors below the x-ray crystallography measurements. While the agreement is not close, there is a clear agreement in the overall pattern of fluctuations. A linear regression gives[5]

$$B^{\mathrm{MD}} \approx 0.192 B^{\mathrm{X}} + 0.594, \tag{8}$$

and the Pearson correlation coefficient between B^{MD} and B^{X} is 0.50. A similar pattern was reported in [2] for the human lysozyme protein.

The differences in the magnitude of the temperature factors can be explained by the sources of error in each estimate. In an MD simulation, the calculated position of each atom is known to machine precision, whereas in an x-ray structure, random noise in the experimental structure factors contributes to errors in the position of each atom and the associated B-factors. Estimated standard deviations (ESDs) of the crambin carbon atom coordinates with B-factors (B_{eq})

[5] Throughout this paper we use the superscript MD to indicate quantities associated with MD simulations and X to indicate quantities calculated from x-ray diffraction experiment data.

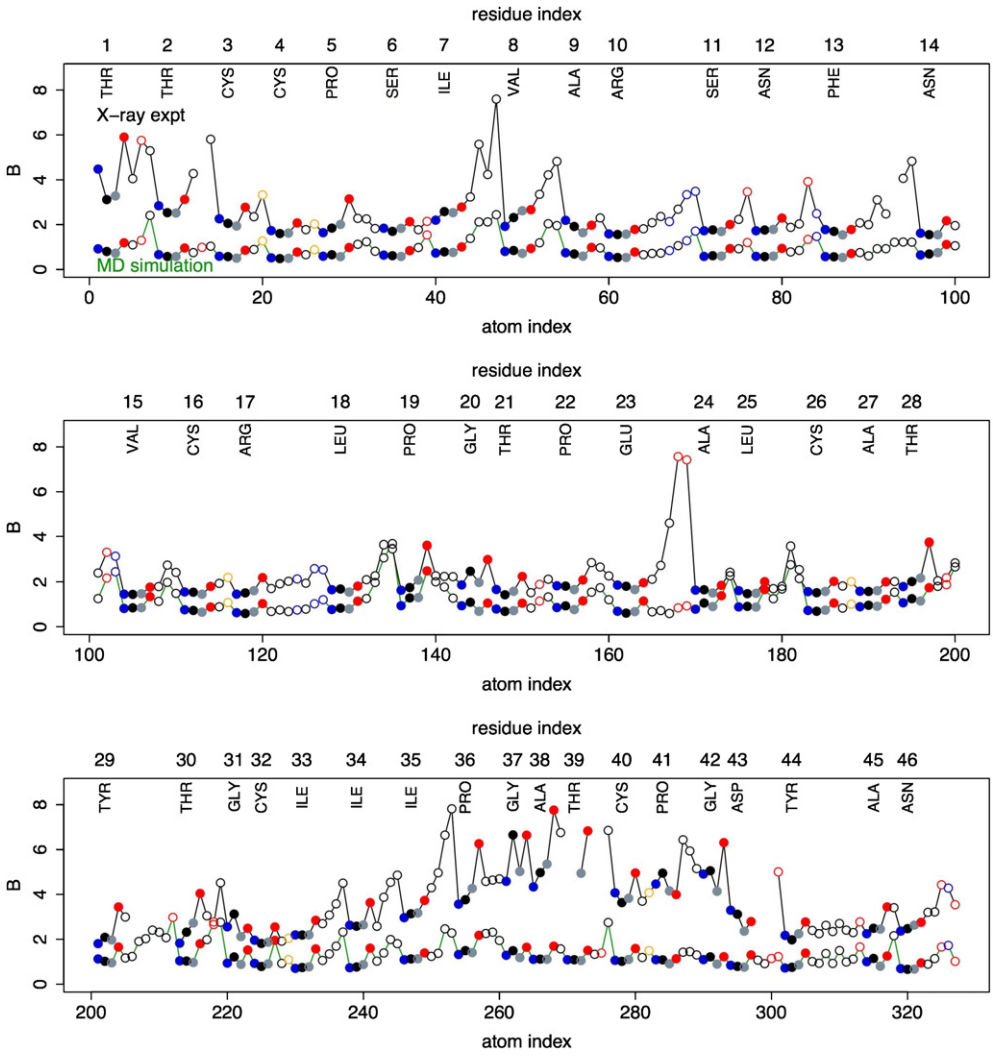

Figure 5. Isotropic temperature factors B^{X} from measured x-ray diffraction data (black line) and B^{MD} calculated from MD simulation (green line) for each of the 327 atoms in crambin. Only B^{X} values calculated from U-matrices with three positive eigenvalues are shown. Dots for each atom are colour coded as follows: blue: nitrogen; black or grey: carbon; red: oxygen; orange: sulfur. Filled-in dots are the backbone (alpha carbon black, carbonyl carbon grey) and the oxygen attached to the carbonyl. Empty circles are side chain atoms.

around 2 Å^2 are about 3.5×10^{-3} Å [9], with ESDs increasing linearly with increasing B_{eq}.

Phonons (lattice vibrations), lattice defects, and internal static disorder of crystals further contribute to B-factors. The x-ray structure represents an ensemble average over all molecules in the crystal, whereas the MD simulation is of one unit protein.

In the present case of crambin, two residues contain different amino acids at a single position: crambin contains proline or leucine at residue position 22 and leucine or isoleucine at residue position 25. While we have chosen the dominant residues for the present MD simulation, the heterogeneity in the crystal will further contribute to differences in x-ray and MD-derived B-factors.

For some atoms the U matrices from x-ray measurements do not have three positive eigenvalues. For these cases, B^{X} as defined by equation (4) is not meaningful and has been omitted from the plot. We believe this may be caused by the atoms in question existing in two well-separated conformations, in which case the true distribution is not well approximated by a trivariate normal. Interestingly, our MD simulation has not registered any of these atoms as being bistable; that is, the Shapiro–Wilk test P-value described in the last section is not less than 0.05 for any of these atoms, but may be found distributed over the whole interval [0, 1].

Conversely, the atom which shows the most pronounced bistability in figure 4, namely the oxygen-γ in residue 30, is one of the few atoms for which B^{MD} is not significantly less than B^{X}. The higher value of B^{MD} is an artefact of the bistablility which does not appear to be present in the experimental data. A similar situation is present to a lesser extent for other atoms exhibiting bistability in figure 4.

4. Anisotropic temperature factors

The isotropic B-factors account for the overall scale of thermal vibrations but contain no information about the shape or orientation of an atom's spatial probability

distribution. Komeiji *et al* [2] have attempted to assess the agreement between orientations of ATFs determined from MD simulations and x-ray crystallography by considering the distribution of acute angles between the respective major axis vectors $\mathbf{a}_1^{\mathrm{MD}}$ and $\mathbf{a}_1^{\mathrm{X}}$ (as defined in figure 1). Contrary to their expectations, they found the expectation value over all atoms of this angle to be slightly higher than that predicted by an isotropic null distribution and concluded that the x-ray and MD results showed no correlation[6].

The use of the angle between $\mathbf{a}_1^{\mathrm{MD}}$ and $\mathbf{a}_1^{\mathrm{X}}$ as a measure of the distance between two ATFs has a number of shortcomings. Firstly there is no measure of agreement between the shapes of the respective UVSs. One would like to know if U^{MD} and U^{X} both corresponded to oblate or prolate distributions as well as knowing their relative orientations. Secondly, spheroids which are close to spherical or having axial symmetry with $\sigma_1 \approx \sigma_2$ may wrongly register as being a bad match as the major axis $\mathbf{a}_1$ of such spheroids is strongly sensitive to small variations in the atomic data. Here we propose an alternative measure of agreement between two independently obtained ATF estimates, which overcomes these difficulties.

4.1. Distance measure between ATFs: mathematical background

Since we are only interested in the shape and orientation of ATFs, having already dealt with the overall magnitude of thermal vibrations in section 3, we define scaled ATFs by

$$\hat{U}^{\mathrm{MD}} = \frac{U^{\mathrm{MD}}}{\operatorname{tr} U^{\mathrm{MD}}}, \qquad \hat{U}^{\mathrm{X}} = \frac{U^{\mathrm{X}}}{\operatorname{tr} U^{\mathrm{X}}}. \tag{9}$$

We propose the distance measure

$$\Delta = \operatorname{tr}[(\hat{U}^{\mathrm{MD}} - \hat{U}^{\mathrm{X}})^2]. \tag{10}$$

This measure is non-negative, and only takes the value zero if the scaled UVSs $\mathcal{S}^{\mathrm{MD}}$ and $\mathcal{S}^{\mathrm{X}}$ (corresponding to MD simulations and x-ray diffraction experiment data, respectively) agree in both shape and orientation.

Without loss of generality, for a given atom we can choose $\mathbf{a}_i^{\mathrm{MD}}$ as defined in figure 1 to be our coordinate axes, so that the scaled ATFs take the form

$$\hat{U}^{\mathrm{MD}} = \begin{pmatrix} \xi_1 & & \\ & \xi_2 & \\ & & \xi_3 \end{pmatrix}, \quad \hat{U}^{\mathrm{X}} = R \begin{pmatrix} \eta_1 & & \\ & \eta_2 & \\ & & \eta_3 \end{pmatrix} R^{\mathrm{T}}, \tag{11}$$

where $R \in SO(3)$ and

$$\sum_{i=1}^{3} \xi_i = \sum_{i=1}^{3} \eta_i = 1, \tag{12}$$

$$\xi_1 \geqslant \xi_2 \geqslant \xi_3 \geqslant 0, \qquad \eta_1 \geqslant \eta_2 \geqslant \eta_3 \geqslant 0. \tag{13}$$

[6] For the human lysozyme protein, Komeiji *et al* obtained $\theta = 59^\circ$ for the average over all atoms of the angle between $\mathbf{a}_1^{\mathrm{MD}}$ and $\mathbf{a}_1^{\mathrm{X}}$, compared with the expected value from a 'null' isotropic distribution of $\hat{\theta} = 57.3^\circ$. For the crambin molecule data, for the equivalent average we obtain a slightly improved value of $\theta = 52.0^\circ$ using the Komeiji *et al* test.

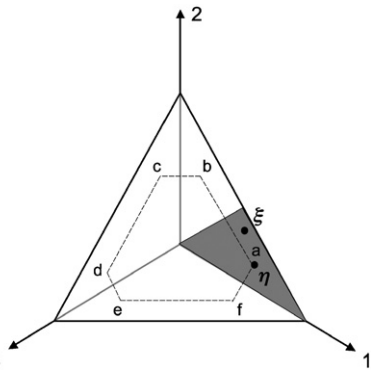

Figure 6. Isometric projection of the 2-simplex, i.e., the equilateral triangle with vertices $(1, 0, 0)^{\mathrm{T}}$, $(0, 1, 0)^{\mathrm{T}}$ and $(0, 0, 1)^{\mathrm{T}}$. The points $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3)^{\mathrm{T}}$ and $\eta = (\eta_1, \eta_2, \eta_3)^{\mathrm{T}}$ are constrained to lie within the shaded subset of the simplex. The point $M\eta$ lies within the region bounded by the hexagon $abcdefa$, the vertices of which are obtained by permuting the three components of $\boldsymbol{\eta}$.

A straightforward calculation then gives

$$\Delta = \sum_{i=1}^{3} \left(\xi_i^2 + \eta_i^2\right) - 2 \sum_{i,j=1}^{3} (R_{ij})^2 \xi_i \eta_j. \tag{14}$$

If $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3)^{\mathrm{T}}$ and $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3)^{\mathrm{T}}$ are held fixed and R is allowed to vary, that is if the shapes of $\mathcal{S}^{\mathrm{MD}}$ and $\mathcal{S}^{\mathrm{X}}$ are held fixed but their relative orientation allowed to vary, we show below that the distance measure Δ ranges over a finite interval $\Delta_{\min} \leqslant \Delta \leqslant \Delta_{\max}$ where

$$\Delta_{\min} = \sum_{i=1}^{3} (\xi_i - \eta_i)^2 \tag{15}$$

is realized when the principal axes are aligned in their correct order, i.e. $\mathbf{a}_i^{\mathrm{MD}} = \mathbf{a}_i^{\mathrm{X}}$, $i = 1, 2, 3$, and

$$\Delta_{\max} = (\xi_1 - \eta_3)^2 + (\xi_2 - \eta_2)^2 + (\xi_3 - \eta_1)^2 \tag{16}$$

is realized when the principal axes are aligned in reverse order, i.e. $\mathbf{a}_1^{\mathrm{MD}} = \mathbf{a}_3^{\mathrm{X}}$, $\mathbf{a}_2^{\mathrm{MD}} = \mathbf{a}_2^{\mathrm{X}}$ and $\mathbf{a}_3^{\mathrm{MD}} = -\mathbf{a}_1^{\mathrm{X}}$.

To understand these limiting values, first note that the restriction equations (12) and (13) constrain the points $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ to lie within the shaded subset of the two-dimensional simplex shown in figure 6. To find extremum values of equation (14), it is sufficient to find extrema of $\Phi = \sum_{i,j} \xi_i M_{ij} \eta_j$, where $M_{ij} = (R_{ij})^2$. Since R is real orthogonal, we have $\sum_i M_{ij} = \sum_j M_{ij} = 1$ and $M_{ij} \geqslant 0$. It follows that $\sum_i (M\boldsymbol{\eta})_i = \sum_{i,j} M_{ij} \eta_j = \sum_j \eta_j = 1$ and $(M\boldsymbol{\eta})_i \geqslant 0$, so the vector $M\boldsymbol{\eta}$ also terminates on the simplex, though not necessarily within the shaded region.

We next show that, for given $\boldsymbol{\eta}$, the point $M\boldsymbol{\eta}$ must lie either within or on the boundary of the region bounded by the hexagon $abcdefa$ in figure 6. The first component of $M\boldsymbol{\eta}$ satisfies

$$\begin{aligned}(M\boldsymbol{\eta})_1 &= M_{11}\eta_1 + M_{12}\eta_2 + M_{13}\eta_3 \\ &\leqslant (M_{11} + M_{12} + M_{13})\eta_1 = \eta_1,\end{aligned}$$

implying that $M\boldsymbol{\eta}$ lies at a point to the left of, or on, the edge af. Similarly, the third component satisfies $(M\boldsymbol{\eta})_3 \geqslant \eta_3$, implying that $M\boldsymbol{\eta}$ lies at a point to the left of, or on, the edge

ab. By writing $M\boldsymbol{\eta} = (MS^2)(S\boldsymbol{\eta}) = (MS)(S^2\boldsymbol{\eta})$ where the matrix

$$S = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

cyclicly permutes the components of any three vectors, one produces similar inequalities corresponding to the corners c and e. These inequalities restrict $M\boldsymbol{\eta}$ to be below and to the left of bcd and above and to the right of def, respectively. Thus $M\boldsymbol{\eta}$ is restricted to be within or on the boundary of the hexagon as the matrix R ranges over all possible elements of $SO(3)$. By choosing the matrix R to be successive rotation matrices about each of the three coordinate axes, it is straightforward to check that the entire boundary of the hexagon is visited, and that the six vertices correspond to rotations affecting the six possible permutations of the three axes.

Finally, we note that $\Phi = \sum_i \xi_i (M\boldsymbol{\eta})_i$ is the projection of $M\boldsymbol{\eta}$ onto the vector $\boldsymbol{\xi}$, which, for any fixed $\boldsymbol{\xi}$ in the shaded region, reaches its maximum at the furthermost right point in the hexagon a, and its minimum value at the furthermost left point d. These points correspond to $M\boldsymbol{\eta} = \boldsymbol{\eta}$ and $M\boldsymbol{\eta} = (\eta_3, \eta_2, \eta_1)^{\mathrm{T}}$, respectively. Direct substitution into equation (14) then confirms equations (15) and (16).

In order to assess the rotational alignment of the spheroids $\mathcal{S}^{\mathrm{MD}}$ and $\mathcal{S}^{\mathrm{X}}$ for each atom, we construct a statistical distribution for the realized value of $\Delta \in [\Delta_{\min}, \Delta_{\min}]$, based on a null hypothesis assumption of spatial isotropy. For this we use the invariant Haar measure for the group $SO(3)$ [10]. Intuitively, this is a distribution designed so that the probability of rotating an object to a given final orientation is independent of the starting orientation. Any rotation R can be parameterized by three Euler angles $0 \leqslant \theta \leqslant \pi, 0 \leqslant \phi_1, \phi_2 \leqslant 2\pi$, as

$$R = \begin{pmatrix} c_1c_2 - Cs_1s_2 & -c_1s_2 - Cs_1c_2 & s_1S \\ s_1c_2 - Cc_1s_2 & -s_1s_2 + Cc_1c_2 & -c_1S \\ s_2S & c_2S & C \end{pmatrix}, \qquad (17)$$

where $C = \cos\theta$, $S = \sin\theta$, $c_k = \cos\phi_k$ and $s_k = \sin\phi_k$ for $k = 1, 2$. The invariant Haar measure is then

$$\mathrm{d}R = \frac{1}{8\pi^2} \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi_1 \, \mathrm{d}\phi_2. \qquad (18)$$

Treating the Euler angles as random variables and $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ as fixed, the invariant Haar measure induces a probability distribution on Δ via equation (14). The mean of the distribution can be found by integration with respect to the Haar measure and is given by

$$\Delta_{\mathrm{mean}} = \sum_{i=1}^{3} \left[\left(\xi_i - \frac{1}{3} \right)^2 + \left(\eta_i - \frac{1}{3} \right)^2 \right]. \qquad (19)$$

We give in table 1 some values of $\Delta_{\min}$, $\Delta_{\max}$ and Δ_{mean} for the extreme cases of UVSs, namely a needle ($\boldsymbol{\xi} = (1, 0, 0)^{\mathrm{T}}$), disc $\left(\boldsymbol{\xi} = \left(\frac{1}{2}, \frac{1}{2}, 0\right)^{\mathrm{T}}\right)$ and sphere $\left(\boldsymbol{\xi} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)^{\mathrm{T}}\right)$. Note that the above procedure defines a four-parameter family of probability distributions parameterized by the locations of the points $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ in the shaded region of figure 6.

Table 1. Values of $\Delta_{\min}$, $\Delta_{\max}$ and Δ_{mean} for extreme cases of UVSs: N = needle ($\boldsymbol{\xi} = (1, 0, 0)^{\mathrm{T}}$), D = disc ($\boldsymbol{\xi} = (\frac{1}{2}, \frac{1}{2}, 0)^{\mathrm{T}}$) and S = sphere ($\boldsymbol{\xi} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})^{\mathrm{T}}$).

$\mathcal{S}^{\mathrm{MD}}$–$\mathcal{S}^{\mathrm{X}}$:	N–N	D–D	N–D	N–S	D–S	S–S
$\Delta_{\min}$	0	0	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{1}{6}$	0
$\Delta_{\max}$	2	$\frac{1}{2}$	$\frac{3}{2}$	$\frac{2}{3}$	$\frac{1}{6}$	0
Δ_{mean}	$\frac{4}{3}$	$\frac{1}{3}$	$\frac{5}{6}$	$\frac{2}{3}$	$\frac{1}{6}$	0

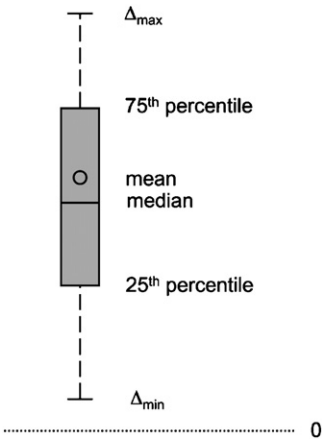

Figure 7. Boxplot showing percentiles and mean of Δ for fixed spheroid shape parameters $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ (defined in equation (11)) as the relative orientation between the spheroids $\mathcal{S}^{\mathrm{MD}}$ and $\mathcal{S}^{\mathrm{X}}$ is varied.

4.2. Graphical representation of Δ

Information about the shape, relative orientation and alignment of the UVSs $\mathcal{S}^{\mathrm{MD}}$ and $\mathcal{S}^{\mathrm{X}}$ for each atom can be presented compactly by superimposing the realized value of Δ on a 'box-and-whisker' plot [11] of the isotropic probability distribution over relative orientations as described above. A typical boxplot showing the median, extremes and quartiles of the distribution is shown in figure 7. In the following numerical analysis of the crambin data, a sample of 10 000 sets of Euler angles was first generated from the Haar measure by randomly generating points $\phi_1 = 2\pi X_1, \phi_2 = 2\pi X_2, \theta = \arccos(2X_3 - 1)$ where the X_i are uniformly distributed random variables on the interval $[0, 1]$. For each atom, a corresponding sample of Δ values was then calculated using equations (14) and (17), from which a boxplot can be drawn using the R programming language function `boxplot(, range = 0)`.

We now give a qualitative description of the application of this procedure to the crambin molecule. The green squares in figure 8 are a plot of Δ for atoms within three of the residues in the in the crambin molecule, residue number 10 (arginine), number 5 (proline) and number 36 (proline). These points are superimposed on boxplots as described above. The boxplots contain information about the shapes of the UVSs $\mathcal{S}^{\mathrm{MD}}$ and $\mathcal{S}^{\mathrm{X}}$ corresponding to the MD simulation and x-ray diffraction experiments, respectively, while the position of the realized value of Δ within the boxplots contains information about their relative alignment.

To read figure 8 one applies the following rules. Rules (i) to (iii) follow from equations (15) and (16).

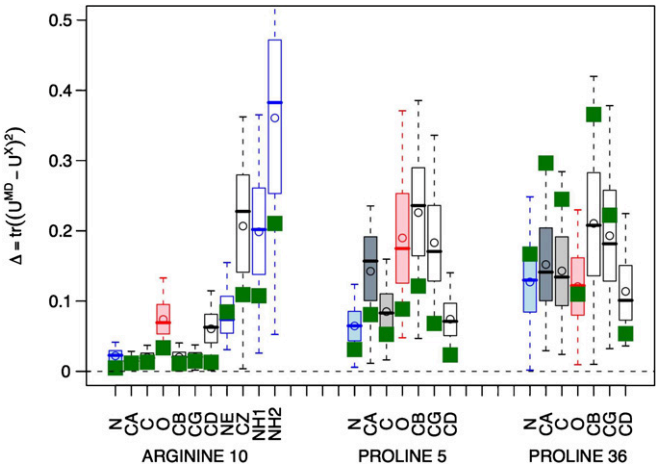

Figure 8. The ATF distance measure Δ (green squares) for atoms within three of the residues in the crambin molecule, superimposed on boxplots of the range of values that Δ can take as the relative orientation between the UVSs from the MD simulation and x-ray diffraction measurements is varied. Atoms are colour coded as follows: blue: nitrogen; black: carbon; red: oxygen. Filled-in boxes are the backbone (alpha carbon dark grey, carbonyl carbon light grey) and the oxygen attached to the carbonyl. Empty boxes are side chain atoms.

(i) If the end of the bottom whisker is close to the $\Delta = 0$ line, $\mathcal{S}^{\mathrm{MD}}$ and $\mathcal{S}^{\mathrm{X}}$ have a similar shape.

(ii) If the end of the top whisker is high, at least one of the two spheroids is either very oblate or very prolate (i.e. very far from being spherical).

(iii) If the distance between the two whisker ends is small, then at least one of the spheroids is close to being spherical.

(iv) The green square gives the actual value of Δ which is realized, and must lie within the box-and-whisker plot. The smaller the value of Δ the better the overall agreement between $\mathcal{S}^{\mathrm{MD}}$ and $\mathcal{S}^{\mathrm{X}}$.

(v) A green square at the end of the bottom whisker is the best possible alignment given the two shapes $\mathcal{S}^{\mathrm{MD}}$ and $\mathcal{S}^{\mathrm{X}}$, and is realized when the three principal axes of the $\mathcal{S}^{\mathrm{MD}}$ align with the axes of $\mathcal{S}^{\mathrm{X}}$ in the same order, i.e. the largest $\mathcal{S}^{\mathrm{MD}}$ axis with the largest $\mathcal{S}^{\mathrm{X}}$ axis and the smallest $\mathcal{S}^{\mathrm{MD}}$ axis with the smallest $\mathcal{S}^{\mathrm{X}}$ axis.

(vi) A green square at the end of the top whisker is the worst possible alignment of $\mathcal{S}^{\mathrm{MD}}$ and $\mathcal{S}^{\mathrm{X}}$ given their shapes, i.e. the largest axis of $\mathcal{S}^{\mathrm{MD}}$ with the smallest axis of $\mathcal{S}^{\mathrm{MD}}$ and *vice versa*.

The absolute position of the green square is a measure of the goodness of fit generally. For instance, if the boxplot is small and low down, both $\mathcal{S}^{\mathrm{MD}}$ and $\mathcal{S}^{\mathrm{X}}$ are close to spherical, and it is not important where within the boxplot the green square is. If the boxplot is elongated but the bottom whisker is close to zero, both of the spheroids are highly asymmetric, rotational alignment is important, and so the match is only good if the green square is in the bottom whisker. If the end of the bottom whisker is clearly above zero, the green square must perforce also be above zero, reflecting the fact that no orientation of $\mathcal{S}^{\mathrm{MD}}$ and $\mathcal{S}^{\mathrm{X}}$ is a close alignment.

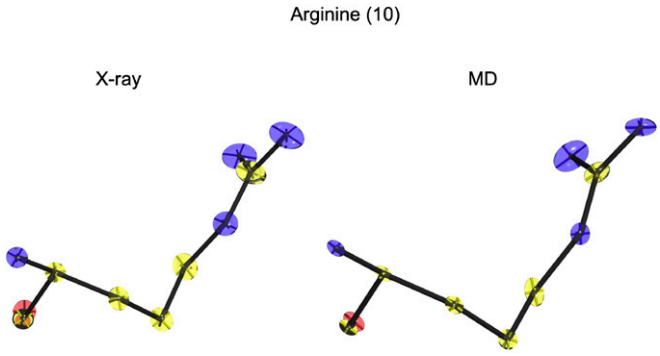

Figure 9. Arginine (residue number 10) showing UVSs. Atoms are colour coded as follows: blue: nitrogen; yellow: carbon; red: oxygen. The backbone is towards the left.

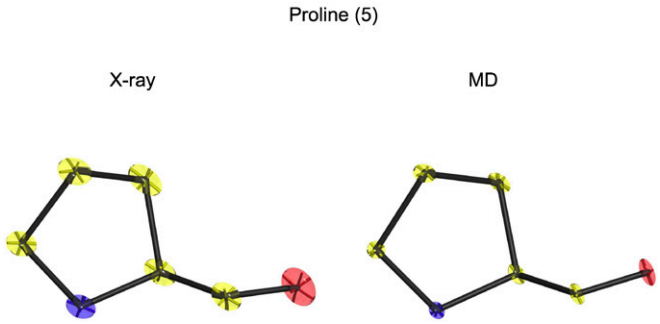

Figure 10. Proline (residue number 5) showing UVSs. Atoms are colour coded as in figure 9. The backbone runs across the bottom.

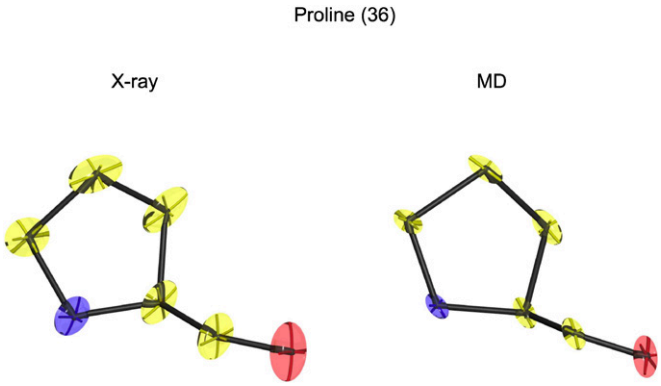

Figure 11. Proline (residue number 36) showing UVSs. Atoms are colour coded as in figure 9. The backbone runs across the bottom.

5. Crambin molecule analysis

Figures 9 to 11 show the arginine and two proline residues for which Δ values are plotted in figure 8. In general, the MD calculation underestimates the isotropic temperature factors as determined by the x-ray diffraction experiment (see figure 5), and this is evident from the relative sizes of UVSs in these images.

Regarding the ATFs, Δ values in figure 8 for the arginine are small and the boxplot whiskers close together for atoms on or close to the backbone, indicating that both the x-ray and MD UVSs are close to spherical. An exception is the oxygen, for which the shape has been less well estimated

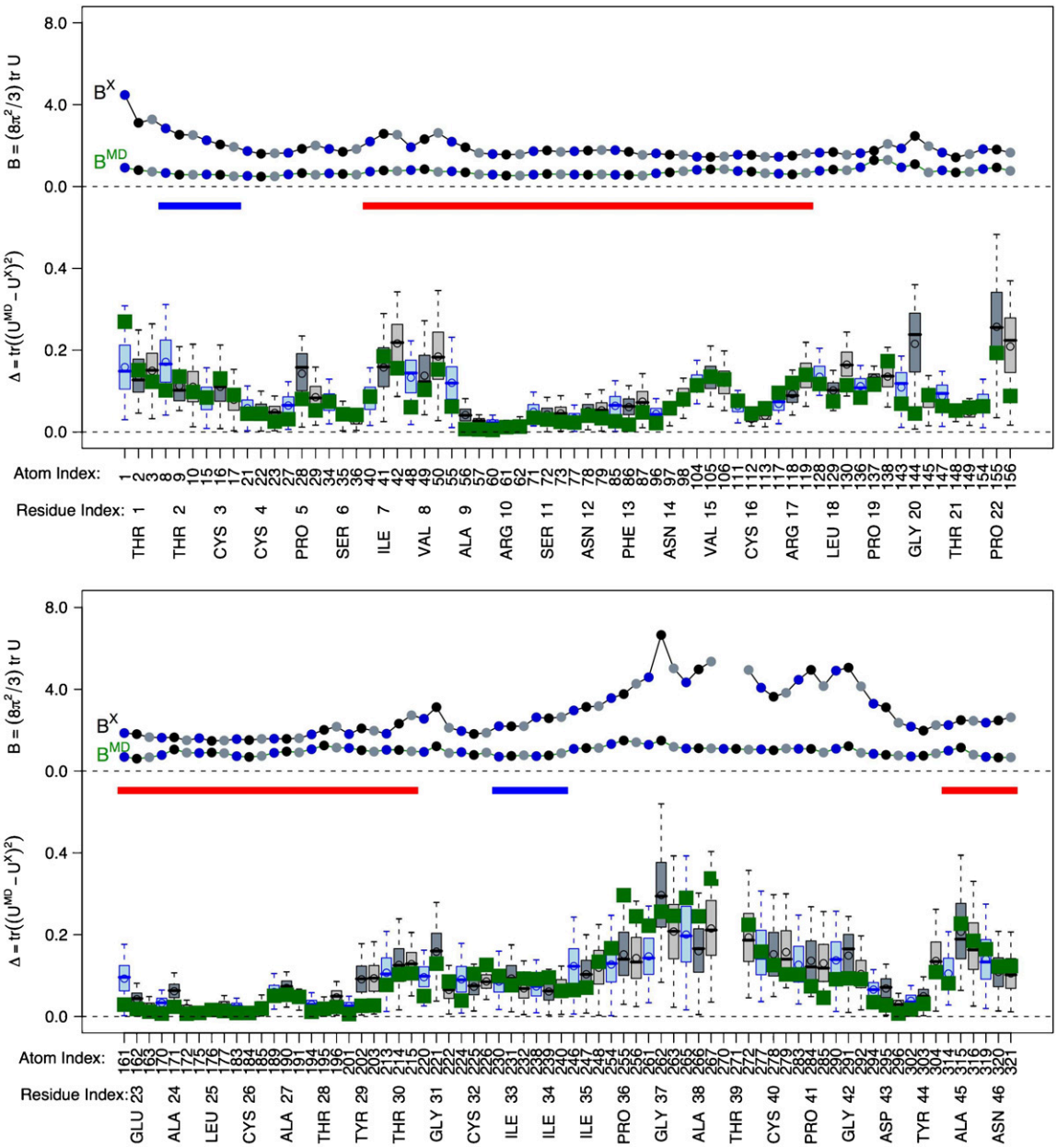

Figure 12. Measured and simulated isotropic temperature factors B^{X} and B^{MD}, and the ATF distance measure Δ for atoms within the backbone of the crambin molecule. Colour coding of atoms is as in figure 7. The red and blue horizontal bars indicate the α-helices and β-sheets, respectively, as shown in figure 2.

(large Δ_{min}), but the orientation correctly determined. The carbon-ζ and nitrogen-η_1 and η_2 at the extremity of the side chain exhibit elongated spheroids, the orientations of which have been reasonably well determined. These features are evident in figure 9.

For the two proline residues in figure 8, the elongated boxplots indicate spheroids which are far from spherical. In both cases the shape of the backbone nitrogen spheroid has been well approximated. The orientation of the spheroids has been well determined by the MD simulation for all atoms in residue number 5, but not for atoms in residue number 36. These features are evident in figures 10 and 11.

Figure 12 shows plots of isotropic temperature factors B^{X} and B^{MD}, and anisotropic Δ values and associated boxplots for all atoms along the crambin backbone. An obvious pattern which emerges is that peaks in B^{X} are generally aligned with longer Δ-boxplots. That is, the largest thermal oscillations tend to be the most anisotropic. For these backbone atoms, the MD simulation fails to reproduce the magnitude of the peaks in B^{X}, and completely misses the peaks at the right-hand end of the largest α-helix in figure 2 (residues 7 and 8) and the unconstrained region (residues 37 to 42). On the other hand, the MD simulation performs better at reproducing peaks in B^{X} for non-backbone atoms (see figure 5).

The smallest thermal oscillations (residues 10 to 17 and 23 to 30) occur where the two main α-helices run parallel to each other. The Δ-boxplots indicate that thermal oscillations are close to spherical in these regions, and this has been

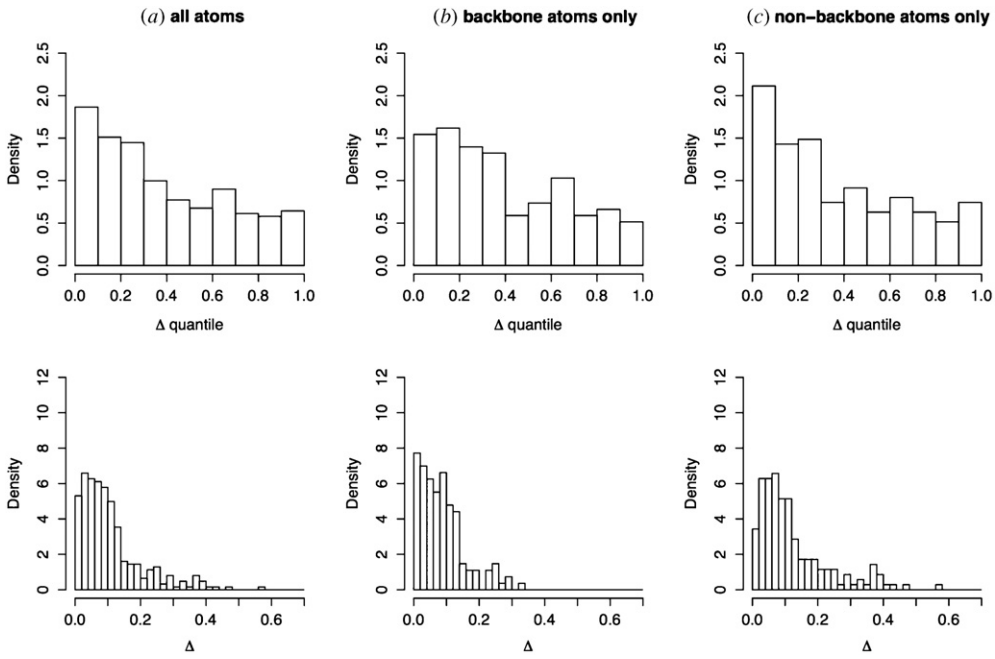

Figure 13. Histograms of the quantiles of the realized value of Δ within the isotropic distribution and of Δ for (*a*) all atoms within the crambin molecule, (*b*) the backbone carbon and nitrogen atoms and (*c*) non-backbone atoms.

successfully reproduced by the MD simulation. The same is true to a lesser extent for the β-sheet (residues 2 to 3 and 33 to 34), though the MD simulation has somewhat underestimated the magnitude of the oscillations. At the right-hand end of the lower α-helix in figure 2 (residues 7 and 8) and the hairpin between the α-helices (residues 19 to 22) the Δ-boxplots show higher isotropy, and the Δ values themselves indicate that the MD simulation has done a better than random job of reproducing the correct orientation.

The MD simulation has performed worst in the region from residues 36 to 42. Here the oscillations are anisotropic, and while the MD simulation has generally performed well in predicting the shape of the UVS (the bottom whisker of the boxplots is close to zero), the orientation is incorrect. The proline in figure 11 is in this region.

Finally, we show in figure 13 histograms of quantiles of the realized values of Δ within the isotropic distribution represented by the boxplots in figure 8. A histogram skewed towards lower quantiles indicates better agreement between the rotational alignment of ATFs determined from MD simulations and from x-ray diffraction measurements. Also shown are histograms of Δ itself. These histograms illustrate the importance of considering the distance function Δ, rather than the rotational alignment of the principal axes of the UVSs. The quantile histograms considered on their own would suggest surprisingly better alignment between the MD simulation and experiment for non-backbone atoms than backbone atoms. However, the UVSs for backbone atoms are on average closer to spherical than for non-backbone atoms (see figure 8 for instance), and for an almost spherical UVS absolute agreement between principal axes is of little importance. More significant are the histograms of Δ, which indicate that the MD simulation does a better job of capturing ATFs for backbone atoms than for non-backbone atoms, as expected. This is particularly evident in the tail of the distribution: the worst performing atoms are almost all non-backbone atoms.

6. Conclusions and outlook

We have developed a compact graphical technique for assessing the accuracy of ATFs determined by MD simulations against experimental data reported in PDB files. The technique indicates in a single plot the overall accuracy of a calculated ATF, the extent of anisotropy and accuracy with which the shape and orientation of the spheroid at unit variance of the spatial distribution have been captured. We are unaware of any adequate way of comparing experimental and theoretical data on ATFs for entire macromolecules in a simple, compact format, prior to this work. Plots such as figure 12 give easily digestible information about the shape, orientation and magnitude of anisotropic temperature factors of large parts of a macromolecule at a single glance. Without such plots one is reduced to examining three-dimensional renderings such as figures 9–11—a procedure which is not practical for entire proteins.

The technique has been applied to an MD simulation of the crambin protein molecule for which ultra-high-resolution data exist in the PDB. We find that the MD simulation captures well the dynamics of backbone atoms in two closely aligned α-helices whose thermal fluctuations are small and spherically symmetric. In slightly less constrained sections of the backbone such as β-sheets, atomic fluctuations become less isotropic and the performance of the MD simulation worse. The MD simulation performs least well in unconstrained parts of the backbone or within side chains where atomic fluctuations tend to be larger and highly anisotropic. In general, there is no difference between the performance of

the simulation for atoms involved in crystallographic contacts and those not in contact with neighbouring protein molecules.

The differences between the thermal ellipsoids in MD simulation and x-ray structure are largely due to artefacts arising from the limitations of the empirical force field used in the MD simulation. For example, one known limitation is the propensity for simulated peptides to form π-helices in aqueous and lipid environments (e.g. Pak *et al* [12]). This has been shown to be a force-field artefact [13]. Despite these limitations, our analysis indicates that the anisotropic motions are reproduced surprisingly well by MD. Pathological cases are generally found in loop regions that are less conformationally restricted (as demonstrated by increasing B-factors). In these cases the directionality of the ellipsoids may be more sensitive to local physical environment and more prone to deficiencies in the force field. Conversely, the main-chain residues of an alpha helix, which are restrained by contacts of main-chain atoms to nearby resiudes ($i + 4$ and $i - 4$) are in excellent agreement.

A further limitation of the MD simulation is limited sampling of phase space. The fluctuations in biomolecules occur over a range of time-scales. While individual atomic fluctuations occur on the femtosecond timescale, rigid body motions such as domain hinge-bending and helical motions occur on timescales from 10^{-9} to 10^{0} s. Thus, while the exploration of each atom's local minima might be expected to converge, rarer events, such as an atom jumping from one local minimum to another, may not be completely sampled. This is particularly important for side chains which are observed in the crystal structure to lie in multiple conformations.

The higher B-factors from the x-ray structure versus the MD simulation may be attributed to a number of factors. Lattice vibrations and lattice defects are not accounted for in the simulation. One crucial factor affecting the x-ray data (but not the MD simulation) is radiation damage. Such damage manifests itself as a loss of diffraction intensity and an increase in the temperature factor. Specific forms of damage occur in the form of breakage of disulfide bonds, decarboxylation of aspartate and glutamate, loss of hydroxide groups from tyrosine, and the loss of methylthio groups from methionine [14]. This damage will not only affect the B-factor of the removed atom but will also purturb the dynamics of surrounding atoms through the changed chemical environment.

The techniques introduced in this paper have broader application than the case considered here of comparing x-ray diffraction experiments with molecular dynamics simulations. In our subsequent work we are finding plots analogous to figure 12 to be useful for comparing independent x-ray diffraction experiments on the same protein, or separate molecular dynamics simulations of the same protein. In general, the new methods presented here are potentially applicable to the comparison of any two data sets containing independent measurements or simulations of anisotropic temperature factors for a given biological macromolecule.

Acknowledgment

We thank Alan Welsh for stimulating discussions.

Glossary

Anisotropic temperature factor. The variance–covariance matrix of an atom's displacement within a unit cell of a crystal about its equilibrium position due to thermal oscillations, lattice vibrations and different discrete conformations in different unit cells. It is usually denoted by the symbol U (see equation (7)).

Crambin. A relatively small protein of 46 amino acid residues found in the plant seeds of Abyssinian cabbage. It forms a very stable crystal which enables atomic coordinates to be determined by x-ray crystallography to extremely high precision.

Debye–Waller or isotropic temperature factor. The linear mean square displacement of an atom within a unit cell of a crystal about its equilibrium position due to thermal oscillations, lattice vibrations and different discrete conformations in different unit cells. It is usually denoted by the symbol B and conventionally includes a factor of $8\pi^2/3$ (see equation (1)).

Haar measure. A way to assign a volume element to each point in a Lie group (see '$SO(3)$' below) in such a way that the volume element is invariant with respect to the group action.

Shapiro–Wilk test. A statistical test in which the null hypothesis is that the sample in question is drawn from a normal (or Gaussian) distribution.

$SO(3)$. The 'special orthogonal group' in three dimensions. That is, the set of 3×3 matrices R satisfying $R^{\mathrm{T}}R = RR^{\mathrm{T}} = I$, $\det R = 1$ representing rigid body rotations in three-dimensional Euclidean space. $SO(3)$ is an example of a Lie group, that is, a mathematical structure which is both a differential manifold and a continuous group.

References

[1] Worldwide Protein Data Bank. http://www.wwpdb.org/
[2] Komeiji Y, Harata K, Ueno Y and Uebayasi M 2000 Anisotropic motion within a protein: comparison between x-ray crystallography and molecular dynamics simulation of human lysozyme *JPCE J.* **12** 39–48
[3] Jelsch C, Teeter M M, Lamzin V, Pichon-Pesme V, Blessing R H and Lecomte C 2000 Accurate protein crystallography at ultra-high resolution: valence electron distribution in crambin *Proc. Natl Acad. Sci.* **97** 3171–76
[4] Phillips J C, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel R D, Kale L and Schulten K 2005 Scalable molecular dynamics with NAMD *J. Comput. Chem.* **26** 1781–802
[5] Kale L, Skeel R, Bandarkar M, Brunner R, Gursoy A, Krawetz N, Phillips J, Shinozaki A, Varadarajan K and Schulten K 1999 NAMD2: greater scalability for parallel molecular dynamics *J. Comput. Phys.* **151** 283–312
[6] MacKerell A D *et al* 1998 All-atom empirical potential for molecular modeling and dynamics studies of proteins *J. Phys. Chem.* B **102** 3586–616

[7] Ichiye T and Karpus M 1987 Anisotropy and anharmonicity of atomic fluctuations in proteins: analysis of a molecular dynamics simulation *Proteins: Struct. Funct. Gen.* **2** 236–59
[8] Shapiro S S and Wilk M B 1965 An analysis of variance test for normality (complete samples) *Biometrika* **52** 591–611
[9] Guillot B, Viry L, Guillot R, Lecomte C and Jelsch C 2001 Refinement of proteins at subatomic resolution with MOPRO *J. Appl. Cryst.* **34** 214–23
[10] Gel'fand I M, Minlos R A and Ya Z 1963 *Shapiro Representations of the Rotation and Lorentz Groups and their Applications* (Oxford: Pergamon)
[11] Becker R A, Chambers J M and Wilks A R 1988 *The New S Language: A Programming Environment for Data Analysis and Graphics (Wadsworth and Brooks/Cole Advanced Books and Software)* (Pacific Grove, CA: Wadsworth and Brooks/Cole)
[12] Pak Y, Jang S and Shin S 2002 Prediction of helical peptide folding in an implicit water by a new molecular dynamics scheme with generalized effective potential *J. Chem. Phys.* **116** 6831–5
[13] Feig M, MacKerell A D Jr and Brooks C L III 2002 Force field influence on the observation of π-helical protein structures in molecular dynamics simulations *J. Phys. Chem.* B **107** 2831–6
[14] Burmeister W P 2000 Structural changes in a cryo-cooled protein crystal owing to radiation damage *Acta Cryst.* D **56** 328–41