PROTEINS: Structure, Function, and Genetics 46:345–354 (2002)
DOI 10.1002/prot.10035

Why Protein R-factors Are So Large: A Self-Consistent Analysis

Dennis Vitkup,[1,2,3] **Dagmar Ringe,**[1,4] **Martin Karplus,**[3,5*] **and Gregory A. Petsko**[1,4*]
[1]*Rosenstiel Basic Medical Sciences Research Center, Brandeis University, Waltham, Massachusetts*
[2]*Department of Biology, Program in Biophysics and Structural Biology, Brandeis University, Waltham, Massachusetts*
[3]*Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts*
[4]*Department of Chemistry and Department of Biochemistry, Brandeis University, Waltham, Massachusetts*
[5]*Laboratoire de Chimie Biophysique, Institut le Bel, Universite Louis Pasteur, Strasbourg, France*

ABSTRACT **The R-factor and R-free are commonly used to measure the quality of protein models obtained in X-ray crystallography. Well-refined protein structures usually have R-factors in the range of 20–25%, whereas intrinsic errors in the experimental data are usually around 5%. We use molecular dynamics simulations to perform a self-consistent analysis by which we determine the major factors contributing to large values of protein R-factors. The analysis shows that significant R-factor values can arise from the use of isotropic B-factors to model anisotropic protein motions and from coordinate errors. Even in the absence of coordinate errors, the use of isotropic B-factors can cause the R-factors to be around 10%; for coordinate errors smaller than 0.2 Å, the two errors types make similar contributions. The inaccuracy of the energy function used and multistate protein dynamics are unlikely to make significant contributions to the large R-factors. Proteins 2002;46:345–354.**

INTRODUCTION

The information that can be obtained from a protein crystal structure depends on the accuracy of the structural model. In many cases, an accurate high-resolution structure, when combined with complementary biochemical data, allows one to understand protein function in atomic detail. Moreover, the structure can serve as a starting point for molecular mechanics or quantum calculations, which provide further insight into structure-function relationships. The fundamental importance of structural methods in biology makes it essential to have a clear understanding of the criteria used to gauge the accuracy of the resulting models.

Local geometric errors in a protein X-ray structure can be estimated by use of knowledge-based programs such as PROCHECK.[1] PROCHECK compares the geometry of a protein model with a database of canonical geometries derived from a set of high-resolution protein structures. The overall accuracy of a protein model is usually evaluated by the ability of the model to reproduce experimental X-ray data. In protein crystallography, the experimental data are the Bragg reflection intensities. The crystallographic R-factor[2] is the most widely used parameter to measure the disagreement between the amplitudes of the experimentally measured reflections and the protein model. The crystallographic R-factor is defined as:

$$R = \frac{\Sigma_{h,k,l}\|F_{\mathrm{obs}}(h, k, l)| - |F_{\mathrm{calc}}(h, k, l)\|}{\Sigma_{h,k,l}|F_{\mathrm{obs}}(h, k, l)|} \tag{1}$$

where h, k, and l are integers that denote points of the crystal reciprocal lattice and $|F_{\mathrm{obs}}(h, k, l)|$ and $|F_{\mathrm{calc}}(h, k, l)|$ are amplitudes of measured (observed) and calculated reflections, respectively. The plot of the crystallographic R-factor versus resolution (the Luzzati plot) can be used to estimate the average coordinate error in a protein crystal structure under assumption of a Gaussian error distribution.[3] For protein structures solved and refined at better than 2 Å resolution, the Luzzati plot usually gives estimates of the average coordinate error in the range 0.2-0.3 Å.[4] Multiple independent refinements of proteins also give error estimates in this range. For example, the root-mean-square deviation (RMSD) between two independently determined structures of subtilisin[5] is 0.32 Å; the average RMSD between three independently determined interleukin-1-α structures[6–8] is 0.53 Å (see ref. 15).

Typically, for an initial protein model (obtained by methods of either molecular or isomorphous replacement[4]) the R-factor is as high as 50%. For comparison, an R-factor of a completely random acentric protein model is 59%.[9] To improve the initial model, crystallographic refinement methods are used. Well-refined protein structures usually have R-factors in the range of 15–25%, depending on the resolution and data quality. In contrast, for small organic compounds R-factor values as low as 2–5% are standard.[10] The purpose of the present article is to determine why the R-factors of well-refined proteins are so large.

The refinement process involves optimization of a cost function of the type:

$$E_{\mathrm{cost}} = w \cdot \Sigma_{h,k,l}(|F_{\mathrm{obs}}(h, k, l)| - |F_{\mathrm{calc}}(h, k, l)|)^2 + \Sigma_{\mathrm{atoms}}E_{\mathrm{geom}}(x, y, z) \tag{2}$$

*Correspondence to: Martin Karplus, Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138. E-mail: marci@tammy.harvard.edu or Gregory A. Petsko, Rosenstiel Basic Medical Sciences Research Center, Brandeis University, Waltham, MA 02454-9110. E-mail: petsko@brandeis.edu

Received 16 February 2001; Accepted 28 August 2001

TABLE I. Total Reflection Numbers (Generated for Different Resolution Shells) and Observations/Parameters Ratios for the Crambin and Myoglobin Refinements

Resolution shell	No. of generated reflections (independent observations)	Ratio of observations to independent adjustable parameters for an isotropic refinement	Ratio of observation to independent adjustable parameters for an anisotropic refinement
Crambin			
10–3.0 Å	729	0.56	0.25
10–2.0 Å	1,246	0.95	0.42
10–1.4 Å	6,906	5.3	2.3
10–1.0 Å	18,722	14.3	6.4
Myoglobin			
10–3.0 Å	3,000	0.6	0.27
10–2.0 Å	10,000	2.0	0.88
10–1.4 Å	26,000	5.2	2.2
10–1.0 Å	82,000	16.4	7.2

where the first term is the squared crystallographic residual weighted by the parameter w and the second is a geometric term, which is often chosen to approximate a molecular mechanics energy function. The energetic term is required to obtain a meaningful structure due to the low ratio of observations to parameters in the protein refinements. Table I gives the ratio of the observations to independent adjustable parameters for crambin and myoglobin.

For a typical protein refinement at 2 Å resolution the observations/parameters ratio is usually close to one.[10,11] As a consequence, a refinement based purely on the crystallographic residual could easily lead to overfitting of the experimental data and a highly distorted structure. The energy term biases the protein models toward a structure with a canonical protein geometry and low molecular mechanics energy. The weight factor w in the cost function (Eq. 2) determines the relative contribution of the reflection residual term and the energetic term. High values of the weight factor favor refined structures with a small crystallographic residual but often lead to poor geometry. Alternatively, low values of the weight factor favor refined models with good geometry but a high crystallographic residual (high R-factor). In practice, a value of the weight factor is chosen, which roughly equalizes the contributions of the crystallographic and energy term in the optimized refinement cost function.

To further reduce data overfitting in crystallographic refinement, cross-validation by using the free R-factor (R-free) was introduced.[12] In a free R-factor refinement, a test set of experimental reflections (usually comprising 5–10% of total reflections) is set aside and not used in the optimized cost function. The free R-factor is defined as:

$$R_{\text{free}} = \frac{\Sigma_{h,k,l \in \text{test\_set}} \| F_{\text{obs}}(h, k, l)| - |F_{\text{calc}}(h, k, l)\|}{\Sigma_{h,k,l \in \text{test\_set}} |F_{\text{obs}}(h, k, l)|} \qquad (3)$$

where h, k, and l now belong only to the test set reflection points of the crystal reciprocal lattice. Because test set reflections are not used in the cost function optimization, a significant drop in the free R-factor value provides an indication of the unbiased improvements in the protein model.

Because of the low ratio of observations to parameters in protein refinements, an isotropic, harmonic model is usually used to model protein motion. In an isotropic refinement, a single Debye-Waller factor (B-factor or temperature factor) per atom is used to describe atomic motion.[2,13] The relationship between the temperature factors and the atomic fluctuations is given by the formula (4):

$$B = 8\pi^2 \langle \Delta r^2 \rangle \qquad (4)$$

where **B** is a temperature factor for an atom and $\langle \Delta r^2 \rangle$ is the mean square atomic fluctuation of the atom from its equilibrium (average) position. By contrast, in refinements of small organic crystals it is usually possible to use a full anisotropic model to describe protein motion. In such anisotropic refinement, six parameters per atom are used. These parameters define a fluctuation tensor for every atom.

Possible origins of the relatively high protein R-factors were discussed by Lattman,[14] who pointed out that errors in experimental measurements of the Bragg reflections are unlikely to make the major contribution. Such errors can be estimated on the basis of a crystallographic residual between two reflection data sets collected from the same crystal. This finding suggests that contributions to protein R-factors from errors in crystallographic reflection measurements are usually on the order of 5% or less, indicating that they are unlikely to be the major factor in high protein R-factors.

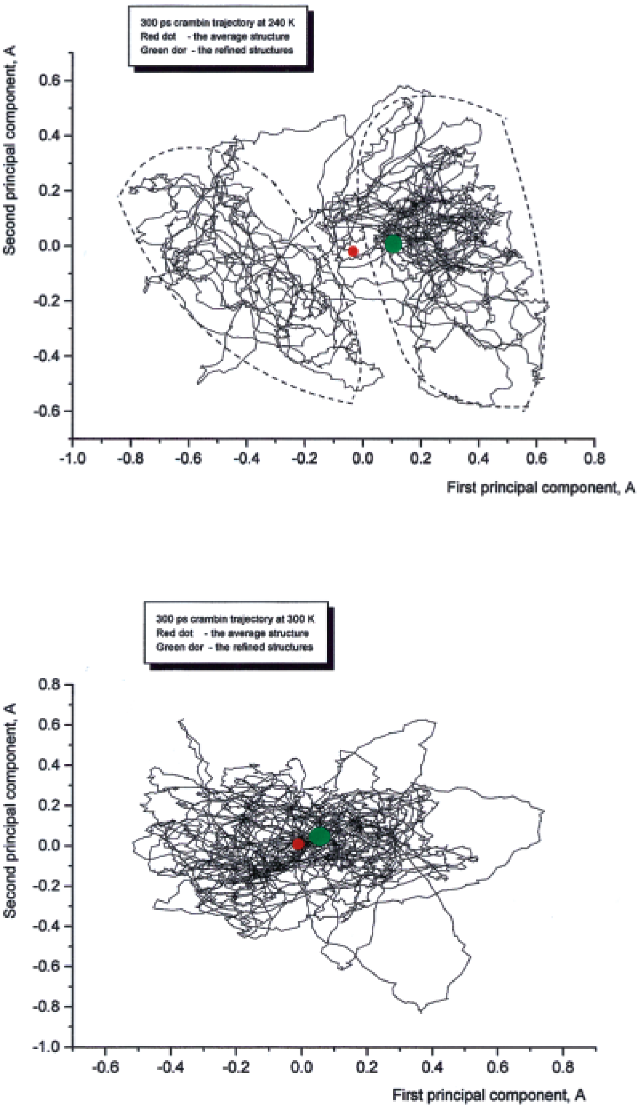

Fig. 1. Projection of the crambin trajectories onto planes defined by two largest principal components. The axes are scaled to represent the root mean square deviations of the structures. The red dot indicates the position of the average trajectory structure, the green dot—the refined structure. The size of the green dot roughly represents the distribution of the structures obtained using slightly different simulated annealing protocols. A. Projection of the crambin trajectory at 240 K; B.) Projection of the crambin trajectory at 300 K.

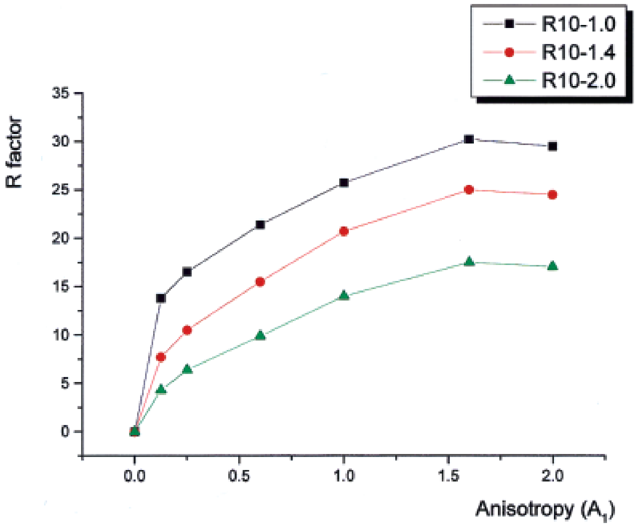

Fig. 2. The R-factor obtained in the isotropic refinements of the harmonic crambin trajectories as a function of the trajectory anisotropy. (for 10-2.0 Å, 10-1.4 Å, 10-1.0 Å resolution shells). The graphs of the refined R-factor and the R-free versus anisotropy were identical for the refinements as the ratio of refined parameters to observations is low.

In the current work, the method of X-ray data generation from molecular dynamics simulations is used to investigate the factors contributing to the relatively large values of refined protein R-factors. An analogous approach was used previously to investigate errors in X-ray protein refinements[11] and of the structures of dissociated CO myoglobin intermediates.[15] X-ray data, generated from molecular dynamic trajectories, are used as an input to commonly used X-ray refinement protocols. Refined results can then be compared directly with known data from the molecular dynamics used to calculate crystallographic reflections. Such an approach is self-consistent and is free from experimental errors and from errors in the molecular dynamics simulations, which allows one to determine the fundamental limitations of the models used to deduce the protein coordinates and motional parameters from experimental data.

RESULTS

Role of Errors in the Energy

To determine whether inaccuracies in the energy term E_{geom} of the crystallographic refinement cost function (Eq. 2) have a significant effect on protein refinements, a self-consistent approach was used. Molecular dynamics trajectories of the small protein crambin (Hendrickson and Teeter, 1981, PDB entry 1crn) were obtained by using CHARMM.[16] Crystallographic reflections in the 10.0–1.0 Å resolution range were generated from the dynamics trajectories as described in Materials and Methods. By using the generated reflections and a refinement cost function with *the same* potential energy term as used in the molecular dynamics simulations, simulated annealing refinements against the generated reflections were performed with the program XPLOR.[17] Thus, differences in the results arise from the refinement procedure and the model on which it is based, rather than from errors in the energy function.

The dependence of the refined protein model on the starting structure, initial velocity assignment in simulated annealing, and general simulated annealing protocol was investigated. Several XPLOR refinement runs, using different starting structures and refinement protocols, all converged to the same structures with backbone RMSD within 0.02 Å from each other. Apparently, the presence of the high-resolution reflection data (up to 1.0 Å resolution) directs refinements to the best possible model independent of the refinement protocols and initial structures.

TABLE II. Characteristics of Crambin Trajectories and the Final R-Factors Obtained in the Refinements of the X-Ray Data Generated From the Trajectories

	Crambin trajectory at 240 K		Crambin trajectory at 300 K		Harmonic crambin trajectory (harm_240 K)	
Average non-hydrogen atoms RMS fluctuation, Å	0.64		0.65		0.64	
Backbone RMSD from the average trajectory structure to the refined structure, Å	0.17		0.09		0.024	
Non-hydrogen RMSD from the average trajectory structure to the refined structure, Å	0.34		0.27		0.029	
Average non-hydrogen atoms anisotropy parameter A_1	0.88		0.71		0.87	
Average non-hydrogen atoms anisotropy parameter A_2	0.15		0.15		0.14	
Resolution range for R-factor calculations	Isotropic R-factor	Isotropic R-free	Isotropic R-factor	Isotropic R-free	Isotropic R-factor	Isotropic R-free
10–2 Å (%)	11.9	13.6	12.2	12.8	11.6	11.9
10–1.4 Å (%)	15.7	17.4	16.3	17.5	16.2	16.8
10–1 Å (%)	20.3	21.6	20.8	21.7	21.2	22.9

The final crystallographic R-factors obtained in refinements of the crambin structure against the reflection data generated from the two typical crambin trajectories (at 240 K and 300 K) are shown in Table II.

Similar protein R-factors and R-frees were obtained in refinements of crystallographic data generated from other crambin trajectories, interleukin-2-β trajectories (D. Vitkup et al., in preparation), and myoglobin trajectories.[15] It is clear from Table II that the R-factor values obtained in the refinements of the crystallographic data generated from the molecular dynamics trajectories are similar to the relatively high R-factors usually obtained in experimental protein refinements,[18] despite the fact that identical energy functions were used in generation of the trajectories and in the refinements.

Role of Anharmonic Multistate Behavior

The crambin trajectories at 240 K and 300 K exemplify trajectories with pronounced multistate and single-state behavior. This is made apparent by a principal component analysis (PCA) of the trajectories.[19,20] Coordinates of C-α backbone atoms of crambin were used to define conformations of the protein in the PCA. Figure 1(a) shows a 300-ps molecular dynamics trajectory of crambin at 240 K in projection onto the two largest principal components. The projection shows that the 240 K crambin trajectory exhibits pronounced multistate behavior. In the trajectory, the protein visits two basins, which are approximately delineated by dashed lines in Figure 1(a). The average structure, calculated from the simulation by averaging the atomic coordinates along the trajectory, is marked by the red dot in Figure 1(a).

The structure obtained in the simulated annealing refinements with the isotropic B factors is marked by the green dot. The refined structure is close to, but somewhat different from, the average structure. The RMSD between the average and the refined structure is 0.17 Å for backbone atoms and 0.34 Å for all non-hydrogen atoms. Figure 1(a) shows that the refined structure lies within one of the basins visited during the dynamics. Analogous results were observed in other crambin trajectories (D. Vitkup, unpublished data): the refined structures are close to the average structures but usually lie within the basin (substate) in which protein spends most of the time during the simulation. The PCA projection of a 300-ps crambin trajectory at 300 K on its two largest principal components is shown in Figure 1(b). Only a single basin is explored in the 300 K trajectory. As a result of the single basin (state) behavior, the refined structure [green dot in Figure 1(b)] is closer to the average structure for the 300 K trajectory [red dot in Figure 1(b)] than for the 240 K trajectory. The RMSD between the average and the refined structure for the 300 K trajectory is 0.09 Å for backbone atoms and 0.27 Å for all non-hydrogen atoms. Despite the multistate dynamics in the 240 K trajectory and single state in the 300 K trajectory, the refined R-factor and R-free values are similar in the two cases (see Table I). Both are within the normal range observed for proteins.

To further analyze the contributions of multistate behavior to the crystallographic protein R-factor, crambin trajectories were generated in which the atomic motion was anisotropic but harmonic within a single state. In the harmonic trajectories, the displacement distribution for each crambin atom around its average position was Gaussian. Details of the generation of the harmonic trajectories are given in Materials and Methods.

Because the intensities of crystallographic reflections depend only on the average electron density in the crystal unit cell (and thus only on the average atomic distributions), the reflections generated from an artificial harmonic trajectory are equivalent to the ones produced by an "imaginary" dynamic trajectory in which the atomic motions are anisotropic but harmonic within a single state. A harmonic "trajectory" (referred to as the harm_240 K trajectory below) was generated in which principal components of motion and amplitudes of fluctuations for each crambin non-hydrogen atom were the same as in the molecular dynamics trajectory at 240 K. The crystallographic reflections were generated from the harm_240 K trajectory in the same way as from the molecular dynamics crambin trajectories. The crambin structure was refined against crystallographic reflections generated from the harm_240 K trajectory by using the simulated annealing refinement method. The refined structure for the harmonic trajectory was almost identical to the average structure for the trajectory. The backbone RMSD between the refined and average structures was 0.024 Å; the non-hydrogen RMS deviation was 0.029 Å. The refined R-factor and R-free for the harm_240 K trajectory are given in Table II. Despite the fact that the harm_240K trajectory was absolutely harmonic, the refined R-factor is about the same as from the 240 K and 300 K trajectories.

Role of Isotropic Approximation to Anisotropic Atomic Motion

After the isotropic refinements of the crambin trajectories using X-PLOR, the crystallographic data generated from the crambin trajectory at 300 K and harmonic crambin trajectory (the harm_240 K trajectory) were refined anisotropically. The anisotropic full-matrix least-square refinements were performed by using program SHELX-93[21] (see Materials and Methods). The final structures obtained in the isotropic refinements with XPLOR were used as a starting model for the SHELX anisotropic refinements. Introduction of anisotropic temperature factors substantially lowered both the R-factors and the R-free factors; the latter confirms the improved quality of the refined protein models. In the anisotropic refinement of the reflection data from the crambin trajectory at 300 K, the R-factor converged at 9.2% and the R-free at 10.0% for all data in the resolution shell 10.0–1.0 Å; the R-factor was 7.1% and the R-free was 8.5% for data in the resolution shell 10.0–2.0 Å. In the refinement of the harmonic crambin trajectory data, the R-factor converged at 4.1% and the R-free at 5.4% for the 10.0–1.0 Å shell; the R-factor was 3.6% and the R-free was 4.5% for the 10.0–2.0 Å shell. Although in the anisotropic refinement of the harmonic trajectory reflections the refined R-factor has not reached zero (which would indicate a perfect model), its value is within that expected from the intrinsic errors in the simulated crystallographic data (see Materials and Methods).

Only small coordinate shifts occurred in the anisotropic refinements relative to the starting structures obtained with isotropic refinement. The non-hydrogen atom RMSDs between the starting and the final refined structures were 0.067 Å for the data generated from the 300 K crambin trajectory and 0.045 Å for the data from the harmonic trajectory. The small RMS coordinate shifts observed in the anisotropic refinements indicate that the major improvement in the R-factors came from the introduction of the anisotropic temperature parameters.

To determine how the lowest R-factors achievable in isotropic refinements are affected by the degree of anisotropic atomic motion, harmonic crambin trajectories with various degrees of anisotropy were generated. The degree of anisotropy in these harmonic trajectories was altered by changing the ratio of atomic fluctuations along the first principal component of the individual atomic motions relative to fluctuations along the other two components. This procedure effectively changes the anisotropic parameter A_1 (see Materials and Methods). The directions of the principal components for the motion of the crambin atoms were the same in all harmonic trajectories and were taken from the molecular dynamics trajectory of crambin at 240 K (see above). The anisotropic parameter A_2 (see Materials and Methods) for all crambin atoms was kept at zero (fluctuations along the second and the third principal components were the same for all atoms), as observed values of the A_2 parameter are small in both molecular dynamic simulations[22] and experiment.[23] The degree of anisotropic motion was varied in the different harmonic trajectories but was identical for all atoms within each trajectory. In the harmonic trajectories, the RMS fluctuations for all non-hydrogen crambin atoms were set equal to 0.75 Å. This is a typical value of atomic RMS fluctuations in proteins at room temperature.[24] Crystallographic reflections were generated from the harmonic trajectories in the usual way (see Materials and Methods). The isotropic refinements of the crystallographic data generated from the harmonic trajectories were performed with XPLOR. Only B factors were refined, and the average structures were used as the coordinate models. For a harmonic trajectory, the average structure (identical to the minimal energy structure) constitutes the best possible refinement model. Consequently, there were no coordinate errors in the refinement of harmonic trajectories. Thus, the only errors came from isotropic refinement of harmonic but anisotropic protein atom motions.

The converged R-factors obtained in the refinements of crystallographic data generated from the harmonic trajectories are plotted as a function of the motional anisotropy (the parameter A_1) of the trajectories in Figure 2. The graphs of the refined R-factor and the R-free versus anisotropy were almost identical for the refinements because the ratio of refined parameters to observations is low (because only B-factors are refined, the parameters/observations ratios are 0.26 for 10–2.0 Å data, 0.047 for 10–1.4 Å data, and 0.017 for 10–1.0 Å data) so that "overfitting" is not possible. The R-factor data for three resolution shells (10–2.0 Å, 10–1.4 Å, 10–1.0 Å are displayed in Figure 2. It is evident from Figure 2 that, even in the absence of coordinate errors, relatively large R-factors (and R-frees) occur because of the inconsistency between the isotropic refinements and the anisotropy of the atomic motion. The graphs of the R-factors versus the motion anisotropy in Figure 2 show that the higher the resolution of the reflection data the larger are the values of the R-factors because of the use of the isotropic approximation. Qualitatively, this is a consequence of the fact that at a higher resolution differences between isotropic and an anisotropic electron distributions are more significant.

Although the data presented in Figure 2 were calculated by using the harmonic trajectories with the principal components of individual atomic motion obtained from the 240 K trajectory, the form of the R-factors versus anisotropy graphs appears to be general. Refinements of trajectories with randomly generated principal components of motion, but the same degree of anisotropy (values of A_1), produced very similar R-factors (D. Vitkup, unpublished).

Role of the Random Coordinate Errors

To complement the study of contributions to the protein R-factors from the use of the isotropic approximation to the atomic motion, it is of interest to consider a system in which atomic motion is isotropic and the contributions are coming exclusively from coordinate errors. To simulate such a system, X-ray reflections were generated from a static crambin structure with preset isotropic B-factors. This approximates reflection data generated from a perfectly harmonic and isotropic simulation. The crambin structure, used to simulate the isotropic reflection data, was then displaced by minimization or brief molecular dynamics, and new reflection data were generated from the displaced structure. The same B-factors were used for all crambin atoms in the original and displaced structures. The protein R-factor between the reflection data generated from the original and displaced structures provides an estimate of R-factor values for a system in which the only source of error is due to the deviation of a structure from the best possible refinement solution.

Figures 3 show three-dimensional graphs of the R-factor between two structures as the function of the RMSD and atomic B-factors. Identical B-factors were used for all crambin atoms in the two structures. The R-factor for each RMS value in Figures 3 was averaged over three independent displacements obtained from short molecular dynamics runs or minimization. Figure 3(a) shows data for 10–1 Å resolution shell, Figure 3(b) for 10–1.4 Å resolution shell, and Figure 3(c) for 10–2 Å resolution shell. As is clear from Figures 3, the surface of R-factor versus coordinate errors is rather steep, and random RMS errors in the range of 0.2–0.3 Å increase R-factors to 20–30%. By contrast, the magnitude of the B-factors has only small effect, with R-factors as a function of coordinate error being slightly larger for smaller B-factor values.

R-Factors Resulting From Combination of Using the Isotropic Approximation and Coordinate Errors

In the refinement of the experimental data for proteins, both coordinate errors and errors due to the isotropic approximation of the atomic motion are usually present. The individual contributions of these errors to the value of the R-factor were investigated in the previous two sections. Here we examine the effect of the presence of both types of errors. The combined contributions of the two error types were estimated by using the following procedure. The final structures obtained in the isotropic refinement of the harmonic and anisotropic crambin trajectories (see above) were displaced by brief molecular dynamics simulations or minimizations. The R-factors were calculated between the reflections generated in the anisotropic trajectories and the displaced structures with isotropic B-factors (Fig. 4). For each anisotropy and value of the coordinate shift, the R-factors shown are the average over 10 independent displacements. Data for 10–1.0 Å resolution shell is shown in Figure 4(a), data for 10–1.4 Å resolution shell in Figure 4(b), and data for 10–2.0 Å resolution shell in Figure 4(c).

Compared with the R-factors in Figure 3 (in which only coordinate errors were considered), the R-factors for trajectories with large anisotropy have significant values even when coordinate errors are near zero. This is a consequence of the fact that the isotropic approximation alone can result in large R-factors (see Role of Isotropic Approximation to Anisotropic Atomic Motion). For coordinate errors smaller than 0.2 Å, the use of the isotropic approximation becomes an important factor keeping values of the R-factors high.

To establish the generality of the results in Figure 4, the R-factors for myoglobin were calculated as a function of anisotropy and RMS displacements. Harmonic myoglobin trajectories were generated in the same way as harmonic crambin trajectories (see above). The principal axes of atomic motion used in generation of harmonic myoglobin trajectories were taken from a 200-ps molecular dynamics trajectory of myoglobin at 300 K generated by CHARMM. The same fluctuation amplitudes were used in generation of the myoglobin harmonic trajectories as in the crambin harmonic trajectories. The R-factor surface plots for myoglobin are compared in Figure 5.

The R-factor plot for resolution shell 10–1.4 Å is shown in red for crambin and in blue for myoglobin. The surface plots of R-factor versus anisotropy and RMS displacements are similar for the two proteins. Moreover, the surfaces are almost identical in the area of usual protein motion anisotropy (anisotropy around 0.7), and typical coordinate errors present in experimental X-ray structures (coordinate errors of 0.2–0.3 Å). The similarity of the R-factor dependence on the degree of motional anisotropy and coordinate errors for proteins as different as crambin and myoglobin suggests that the shape of the plots is a general feature of protein refinement.

DISCUSSION

We have explored the contribution of various aspects of protein motion and refinement models to the observed values of R-factors. This was done by the use of molecular dynamics simulations and their self-consistent analysis that permitted a dissection of the contributions to the R-factor that could not be made by experiments alone. As stated in the introduction, the errors in the measurement of crystallographic reflections, although present, usually do not play the dominant role in keeping protein R-factor values high. With modern high-intensity sources and area detectors, individual reflection intensities are measured quite accurately. The errors of this type were estimated to contribute <5% to the observed R-factors.[14]

The effects connected with solvent modeling and disorder in protein crystals clearly contribute to protein R-factors. It would be possible to extend our self-consistent analysis to investigate these effects by performing solvated simulations in a crystal environment. We note that the R-factors obtained in our study are similar to experimental values, suggesting that solvent modeling and crystal disorder are not the principal contributors to experimental R-factor values.

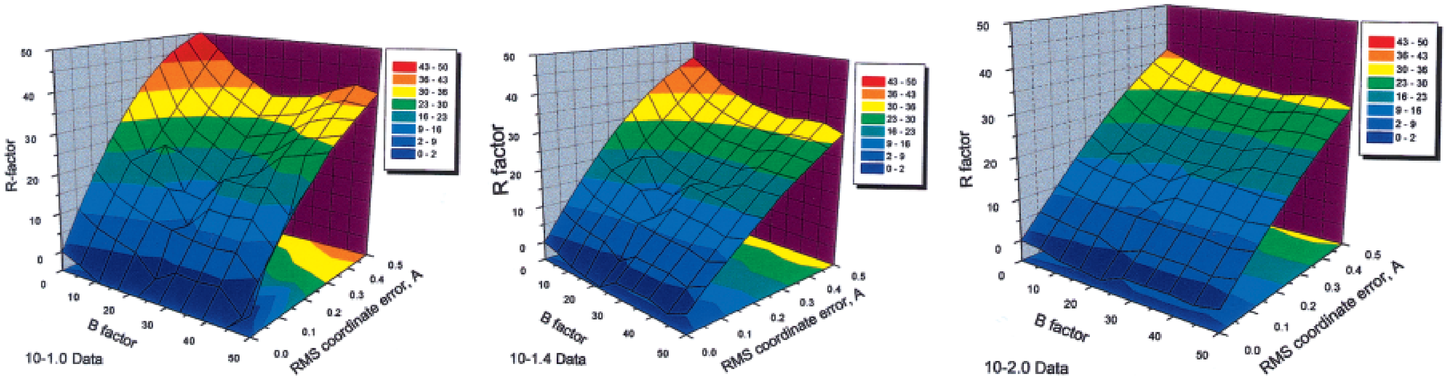

Fig. 3. The R-factor between crystallographic reflections set generated from two crambin structures, with the same B-factors assigned to all non-hydrogen atoms, as a function of all-atom RMS deviation between the structures (coordinate errors) and the value of the B-factors. The displayed value of the R-factor is an average obtained from three independent coordinate displacements. Coordinate displacements were obtained by short molecular dynamics runs or protein minimization. a) R-factors for reflections in 10-1.0 Å resolution shell, b) R-factors for reflections in 10-1.4 Å resolution shell, c) R-factors for reflections in 10-1.0 Å resolution shell.

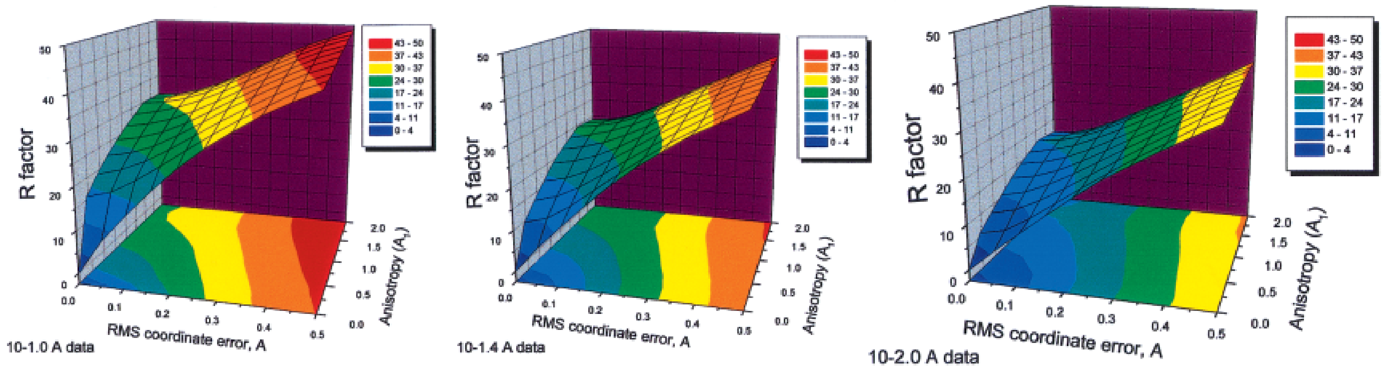

Fig. 4. The R-factor between crystallographic reflection sets generated from the anisotropic harmonic trajectories of crambin and displaced structures (by short molecular dynamics runs or minimizations) obtained in the isotropic refinements of these trajectories. The R-factors are shown as a function of all-atom RMS displacement and anisotropy of the trajectories (see text). a) R-factors for reflections in 10-1.0 Å resolution shell. b) R-factors for reflections in 10-1.4 Å resolution shell. c) R-factors for reflections in 10-1.0 Å resolution shell.

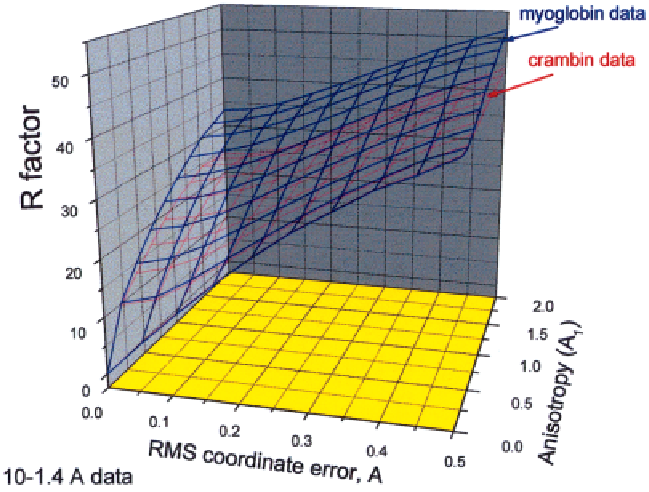

Fig. 5. The graph analogous to Figures 4a–4c, comparing the R-factor surfaces for crambiin and myoglobin.

Inconsistencies between the "real" protein energy function and the functions used as an energetic term in refinement protocols can also be ruled out as the major source of the large protein R-factors. Our results show that even when using the exact energy function in the refinement, the R-factors are not lowered substantially. In a protein refinement, one usually tries to approximate an ensemble of conformations with a single structure. The structure with the lowest potential energy is the best representative of an ensemble only if the energy surface of the protein is harmonic. In reality, protein energy surfaces have multiple minima in the neighborhood of the native state.[25] For such a complex surface, the single best representative of the ensemble in terms of electron density may not be the structure with the lowest potential energy. The results also show that multistate effects in protein dynamics[25] are not a major contributor to the large protein R-factors. The refinements of the crystallographic reflections generated from crambin trajectories with multistate and single-state behavior did not yield significantly different R-factors. Furthermore, refinements of the X-ray data generated from absolutely harmonic but anisotropic crambin trajectories (perfect single-state dynamics) led to R-factors similar to those obtained from refinement of multistate dynamics.

Investigation of the contribution of coordinate errors and the use of isotropic approximation to anisotropic protein motion showed that both can be important in keeping protein R-factors large. The results of anisotropic SHELX refinements, using the reflection data generated from the crambin trajectories, showed that protein R-factors can be substantially reduced if it is possible to refine anisotropic temperature factors without data overfitting. In the anisotropic refinement of the X-ray data generated from the 300 K crambin trajectory small R-factors were achieved (R-factor = 9.2%, and R-free = 10.0% for reflections in the 10.0–1.0 Å resolution shell; R-factor = 7.1%, and R-free = 8.5% for reflections in the 10.0–2.0 Å resolution shell). The R-factors obtained in this way are about half as large as those resulting from the isotropic refinements of the same data. The corresponding reduction of the R-free in the anisotropic refinements indicates significant improvements in the quality of the refined models. These improvements were almost exclusively due to better modeling of anisotropic protein motion, because there were only small coordinate shifts in the course of the SHELX refinements.

Even if the atomic protein motions were completely isotropic (i.e., the use of isotropic approximation were exact), coordinate errors can still cause large protein R-factors. By "coordinate errors" we mean the difference between a given structural model and the best possible refinement solution. Coordinate errors can occur in X-ray structures because, for example, reflection data are not of a high enough resolution to guide refinements to the best possible structural model. Typical RMS coordinate errors in protein structures are estimated at 0.2–0.3 Å.[4] Our calculations show that, even if the protein motion is completely isotropic, coordinate errors in the range of 0.2–0.3 Å result in R-factors of 20–25%.

To understand the interplay between contribution from coordinate errors and the use of the isotropic approximation, combinations of these errors were considered in Figure 4. Graphs of the protein R-factors versus coordinate RMS errors and degree of the motion anisotropy were calculated. The analysis of the resulting surfaces led to the conclusion that for coordinate errors below 0.2 Å and typical protein anisotropy (parameter A_1 around 0.7–0.8), about equal contribution to the experimental R-factors are made by these two factors.

Our conclusions are in accord with recent experimental evidence, for example, results obtained by Harata et al.[23] In this study, the SHELX program was used to refine structures of turkey egg white and human lysozyme anisotropically. In the lysozyme refinements, the introduction of anisotropic temperature factors markedly reduced the protein R-factors. The final R-factors achieved by Harata et al. were 10.4% for turkey egg lysozyme and 11.5% for human lysozyme. These values are to be compared with the isotropic refinement that yielded R-factor of 18% and R-free of 23.5%.

In the future, improvements in experimental data collection and refinement techniques should generally decrease the observed protein R-factors. Methods of cryocrystallography, better crystal growth, and use of high-power synchrotron sources will allow measurements of X-ray data to higher resolution. This would decrease coordinate errors in refined structures. Higher resolution data also should allow for wider use of the anisotropic B-factor approximation, especially when combined with normal mode refinement[26,27] and time-averaged refinement.[28] It is likely that refinement R-factors will be reduced to around 10% by such improvements in the near future, at least for some proteins.

MATERIALS AND METHODS

Molecular Dynamics Simulations

The molecular dynamics simulations of crambin, used for generation of X-ray reflections, were performed with the CHARMM program (version 24a).[16] Two molecular dynamic simulations were performed at 240 K and at 300 K. The simulations were started from an energy minimized crambin crystal structure (Hendrickson and Teeter, 1981, PDB entry 1crn). The crystal structure was first minimized for 500 steps by using the ABNR algorithm. The molecular dynamics simulations included a 10-ps heating stage and a 100-ps equilibration stage, followed by 300 ps of production dynamics. An integration step of 0.001 ps was used. No explicit water molecules were included in the simulations. Coordinate frames were saved every 0.1 ps. The production dynamics portions of the trajectories were used in the simulation of X-ray data.

Potential Energy Function Used in CHARMM and XPLOR

The same energy functions were used to perform the crambin simulations (using CHARMM) and as the energy cost function term in the refinements of X-ray data generated from the simulations (using XPLOR). The polar-hydrogen representation was used, defined by CHARMM 19 parameter and topology files.[29] A switching function was used to truncate the van der Waals and electrostatic interactions over the 6.5–7.5 Å interval. A distance-dependent dielectric was used to screen electrostatic interactions.

PCA of the Molecular Dynamics Trajectories

PCA of the crambin trajectories at 240 K and 300 K was performed by diagonalizing the coordinate covariance matrix.[19,20] The Cartesian coordinates of C-α backbone atoms were used to define a 132 dimensional coordinate space (3N-6 dimensions, where N = 46 for crambin). Before PCA analysis, net rotation and translation of the protein were removed from the trajectory frames by coordinate superposition with the average structure. Projections of the crambin trajectories onto planes defined by the two largest principal components were constructed from 3000 coordinate frames (saved every 0.1 ps from the 300-ps trajectories).

Parameters Characterizing the Atomic Motion Anisotropy

To characterize the anisotropic motion in the present article we use two parameters: A_1 and A_2, introduced previously.[22] Parameter A_1 is formally defined for an atom as:

$$A_1 = \left(\frac{\langle U_x^2\rangle}{1/2*(\langle U_y^2\rangle + \langle U_z^2\rangle)}\right)^{1/2} - 1 \tag{5}$$

where $<U_x{}^2>$, $<U_y{}^2>$, and $<U_z{}^2>$ are the mean-square fluctuations along the principal components of motion for the atom. Specifically, $<U_x{}^2>$ is the mean-square fluctuations along the direction of the first (largest) principal component, $<U_y{}^2>$ is the mean-square fluctuations along the second principal component, and $<U_z{}^2>$ is the mean-square fluctuation along the third principal component. The parameter A_1 approximately indicates how large the fluctuations are along the largest principal component of motion in comparison with fluctuations along the other two components. For absolutely isotropic motion, the parameter A_1 is equal to 0. The larger the value of A_1, the more anisotropic is the motion. Typical average values of the parameter A_1 for proteins are 0.6–0.8.[22] The other parameter that is used to characterize the anisotropy of the atomic motion is A_2; it is formally defined as:

$$A_2 = \left(\frac{\langle U_y^2\rangle}{1/2*(\langle U_y^2\rangle + \langle U_z^2\rangle)}\right)^{1/2} - 1 \tag{6}$$

where $<U_y{}^2>$ and $<U_z{}^2>$ are the mean-square fluctuations along the second and the third principal components of atomic motion. The parameter A_2 indicates how large the atomic motion is along the second largest principal component compared to the motion along the third largest principal component. Typical average values of the parameter A_2 for proteins are 0.1–0.2.[22]

Simulation of Harmonic Crambin Trajectories

The frames of the harmonic crambin "trajectories" were generated by independently displacing each crambin heavy atom along three perpendicular axes. The atomic displacements were done in such a way that in the resulting trajectories a three-dimensional Gaussian distribution was generated for each crambin atom with the displacement axes as the principal components of the distribution. Each harmonic crambin trajectory consisted of 300 frames. In every frame of the harmonic trajectories atoms were displaced relative to the same initial positions constituting the centers for the atomic distributions. Atomic positions in a crambin crystal structure (Hendrickson and Teeter, 1981) were used as the centers for the distributions.

Principal component vectors and fluctuations along these vectors for all heavy atoms were calculated for the crambin trajectory at 240 K by using CHARMM. These principal component vectors and fluctuations were used to generate the harm_240 K trajectory. The same principal component vectors (as in the harm_240 K trajectory) were used in generation of the crambin harmonic trajectories with different motional anisotropies. In these trajectories, the RMS fluctuations were kept the same for all atoms at a value of 0.75 Å. The degree of motion anisotropy, defined by the parameter A_1, was varied simultaneously for all crambin heavy atoms by changing the ratio of fluctuations along the main principal component relative to the other two components.

Simulation of X-ray Data

The program XPLOR was used to generate X-ray data from the molecular dynamics and harmonic trajectories. The X-ray data were simulated by generating X-ray crystallographic reflections from the coordinate frames of the trajectories. Three hundred coordinate frames saved from the molecular dynamics trajectories (every 1 ps) or simulated in the harmonic trajectories were used in the X-ray data generation. No explicit "experimental" noise was added to the generated reflection data. In generation of crystallographic reflections, crambin was assumed to be in a $P2_1$ space group with the following unit cell parameters: a = 40.96, b = 18.65, c = 22.52, α = 90.0, β = 90.77, γ =

90.0 (1crn PDB structure); myoglobin was assumed to be in a $P2_1$ space group, with unit cell parameters: a = 64.10 Å, b = 30.84 Å, c = 34.69 Å, α = 90.0, β = 105.84, γ = 90.0 (1mbc PDB structure). The direct summation method in XPLOR was used for the generation of the crystallographic reflections. No explicit temperature factors were used in generation of the reflections. The reflections generated from the coordinate frames were vector averaged. The vector averaging of crystallographic reflections (averaging of structure factor vectors) corresponds to the assumption that the motion of proteins molecules in different unit cells of a crystal are not correlated.[11] The averaged reflections were used as input data to the refinement protocols.

The convergence errors in the simulated X-ray data were estimated by calculating the R-factors between several harmonic trajectories of the same length (300 frames were generated for each trajectory), atomic fluctuation amplitudes, and principal components of motion for all crambin atoms; the only difference between the trajectories were initial seeds for a random number generator. The R-factors between the reflection data generated from these trajectories were around 5% (for all data in resolution shell 10.0–1.0 Å).

Refinement of the Simulated Data

The isotropic refinements of crambin against generated X-ray reflections were performed by using XPLOR. All reflections generated in the 10–1.0 Å resolution shell were used in the XPLOR refinements. Cartesian simulated annealing from a temperature of 1500 K, followed by successive rounds of minor manual rebuilding, positional and individual B-factor refinements were performed. To ensure convergence of the coordinate simulated annealing refinements to the best possible structural models, isotropic temperature factors (B-factors) of crambin heavy atoms were preset to the values corresponding to their fluctuations in the molecular dynamics simulations (from which the crystallographic data were generated). Throughout the refinements, the value of R-free factor[12] was followed to prevent data overfitting. The refinements of the harmonic trajectories were started from the structures that were centers of atomic distributions (known best possible models), and only B-factor refinements were performed.

The anisotropic refinements were performed by using the program SHELX.[21] The final coordinate models obtained in the isotropic refinements by XPLOR were used as starting models for in the SHELX refinements. Successive rounds of least-squares refinements were performed by SHELX using all reflections in resolution shell 10–1 Å. In the early rounds of the refinements, isotropic temperature factors were used. At the final stage, the full-matrix least-square refinements with anisotropic temperature factors were performed.

REFERENCES

1. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structure. J Appl Crystallogr 1993;26:283–291.
2. Stout GH, Jensen LH. X-ray structure determination, a practical guide. New York: John Wiley & Sons; 1989.
3. Luzzati V. Traitment statisque des erreurs dans la détermination des structures cristallines. Acta Crystallogr 1952;5:802–810.
4. Janin J. Errors in three dimensions. Biochimie 1990;72:705–709.
5. Fitzpatrick PA, Steinmetz ACU, Ringe D, Klibanov AM. Enzyme crystal structure in an neat organic solvent. Proc Natl Acad Sci USA 1993;90:8653–8657.
6. Veerapandian B, et al. Functional implications of interleukin-1b based on the three dimensional structure. Proteins 1992;12:10–23.
7. Priestle JP, Schar HP, Grutter MG. Crystallographic refinement of interleukin 1 beta at 2.0 A resolution. Proc Natl Acad Sci USA 1989;86:9667–9669.
8. Finzel BC, Clancy LL, Holland DR, Muchmore SW, Watenpaugh KD, Einspahr HM. Crystal structure of recombinant human interleukin-1 beta at 2.0 A resolution. J Mol Biol 1989;209:779–791.
9. Wilson AJC. Largest likely values for the reliability index. Acta Crystallogr 1950;3:397–398.
10. Drenth J. Principles of protein X-ray crystallography. New-York: Springer-Verlag; 1994.
11. Kuriyan J, Petsko GA, Levy RM, Karplus M. Effect of anisotropy and anharmonicity on protein crystallographic refinement. J Mol Biol 1986;190:227–254.
12. Brunger AT. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature 1992;355:472–475.
13. Petsko GA, Ringe D. Fluctuations in protein structure from X-ray diffraction. Annu Rev Biophys Bioeng 1984;13:331–371.
14. Lattman EE. Why are protein crystallographic R-value so high? Proteins 1996;25:9–11.
15. Vitkup D, Petsko GA, Karplus MA. Comparison between molecular dynamics and X-ray results for dissociated CO in myoglobin. Nat Struct Biol 1997;4:202–208.
16. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4:187–217.
17. Brunger AT. XPLOR manual Version 3.1. New Haven: Yale University Press; 1992.
18. Kleywegt GJ, Jones TA. Model-building and refinement practice. Methods Enzymol 1997;277:208–230.
19. Jackson EJ. A user's guide to principle components. London: John Wiley & Sons; 1991.
20. Caves LSD, Evanseck JD, Karplus M. Locally accessible conformations of proteins—multiple molecular dynamics simulations of crambin. Protein Sci 1998;7:649–666.
21. Sheldrick GM. SHELX-93. Program for crystal structure refinement. Gottingen: University of Gottingen; 1993.
22. Ichiye T, Karplus M. Anisotropy and anharmonicity of atomic fluctuations in proteins: analysis of a molecular dynamics simulation. Proteins 1987;2:236–259.
23. Harata K, Abe Y, Muraki M. Full-matrix least-squares refinement of lysozymes and analysis of anisotropic thermal motion. Proteins 1998;30:232–243.
24. Brooks CL III, Karplus M, Pettitt BM. Proteins: a theoretical perspective of dynamics, structure, and thermodynamics. New York: John Wiley & Sons; 1988.
25. Elber R, Karplus M. Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. Science 1987;235:318–321.
26. Kidera A, Go N. Normal mode refinement: crystallographic refinement of protein dynamic structure. J Mol Biol 1992;225:457–475.
27. Kidera A, Inaka K, Matsushima M, Go N. Normal mode refinement: crystallographic refinement of protein dynamic structure. II. Application to human lysozyme. J Mol Biol 1992;225:447–486.
28. Gros P, van Gunsteren WF, Hol WGJ. Inclusion of thermal motion in crystallographic structures by restrained molecular dynamics. Science 1990;249:1149–1152.
29. Neria E, Fisher S, Karplus M. Simulation of activation free energies in molecular systems. J Chem Phys 1996;105:1902–1921.