Protein flexibility in solution and in crystals

Peter Eastman
Department of Applied Physics, Stanford University, Stanford, California 94305-4090

Matteo Pellegrini
Molecular Biology Institute, University of California, Los Angeles, Los Angeles, California 90095

Sebastian Doniach
Department of Applied Physics and Department of Physics, Stanford University, Stanford, California 94305-4090

(Received 17 December 1998; accepted 26 February 1999)

Characterizing the inherent flexibility of a protein provides an important link between structure and function. In this article, we examine some of the methods used to determine protein flexibility, and address several unanswered questions relating to them. We perform 4 ns simulations of bovine pancreatic trypsin inhibitor (BPTI) in solution and in a crystal. For comparison, we also calculate atomic fluctuations from room temperature x-ray diffraction data by two different methods: single copy refinement with isotropic B-factors, and constrained multiple copy refinement. We reach the following conclusions: (1) Crystal contacts significantly reduce atomic fluctuations, especially in the flexible loop regions. (2) Center of mass motion in the crystal contributes 0.1–0.2 Å to the rms fluctuations, with little variation between parts of the protein. (3) Isotropic B-factors are an accurate measure of atomic motion in the stable parts of the protein, but significantly underestimate motion of the flexible sidechains. (4) Nanosecond scale simulations can obtain a reasonable sampling of backbone atomic motion in the most stable regions, but are still too short to allow flexible regions to explore their full range of motion. © *1999 American Institute of Physics*.
[S0021-9606(99)50220-1]

I. INTRODUCTION

Proteins are not rigid. This fact is well known, yet surprisingly easy to overlook. The ''native state'' is not a single conformation, but a whole ensemble of conformations which are populated under physiological conditions. In many cases, this flexibility is essential to biological function.[1,2] For example, it allows enzymes to fit themselves around their ligands, and molecular motors to convert chemical energy into mechanical work. Therefore, characterizing the flexibility of a native protein is an issue of considerable interest in relating structure to function. The goal of this article is to examine some of the methods for doing this, and to address some of the unanswered questions relating to them.

Most methods for studying flexibility involve measuring thermal fluctuations. One observes the protein under equilibrium conditions, and measures how much each atom fluctuates about its average position. To the extent that the protein is free to explore its full range of motion, these thermal fluctuations offer a direct measure of the protein's inherent flexibility.[3]

One of the most frequently used methods for doing this is x-ray crystallography.[4] When a model is fit to crystallographic data, the fitting procedure yields a ''B-factor'' for each atom, which describes how widely the electron density is spread out around that atom's location. This is generally interpreted as a measure of the atom's equilibrium fluctuations about its average position, and hence of the protein's local flexibility.[5]

A second experimental method, nuclear magnetic resonance (NMR) spectroscopy, is also widely used for studying equilibrium motion within proteins.[6] This method has an important advantage over crystallography, in that it can provide some information about the time scale on which motions occur. On the other hand, the physical interpretation of NMR data is somewhat ambiguous. The parameters obtained through NMR spectroscopy—generalized order parameters and effective correlation times—can only be assigned a clear physical meaning if one assumes a specific model for the atomic motion.

Molecular dynamics (MD) simulations offer another approach for studying atomic fluctuations.[7] Given a simulation, it is a simple task to calculate the magnitude of each atom's fluctuations about its average position, which can then be directly compared to crystallographic B-factors. Of course, the results will only be meaningful to the extent that the computer model is an accurate representation of the real protein.

Each of the methods discussed above has unique strengths and weaknesses. By combining all three, one can gain a clearer understanding of atomic thermal motion than from any one method alone. Specifically, we examine the following issues:

(1) NMR experiments, and most simulations, study proteins in solution, whereas crystallography deals with proteins in a crystal. Presumably, the contacts with other molecules in the crystal tend to inhibit the protein's motion.[8] Thus, the fluctuations seen in a crystallography experi-

ment may not be a good representation of the protein's inherent flexibility. Just how important this effect is remains an open question.

(2) A molecule's motion can be divided into two components: ''internal motion,'' which consists of the movement of one part of the molecule relative to another; and ''global motion,'' which involves translations of the molecule's center of mass, as well as rotations of the whole molecule about its center of mass. In most cases, only the former is of interest. In NMR and MD, there are straightforward methods for separating the two components and eliminating the effects of global motion from the data. This is not true for crystallography. Even in a crystal, each molecule undergoes some limited amount of global motion, which contributes to the measured fluctuations.[9–11] Once again, the magnitude of this contribution remains an open question.

(3) By calculating an average quantity from a computer simulation, one implicitly assumes that the range of conformations seen by the simulation represent a reasonable approximation of the thermodynamic ensemble. It is not at all clear, however, that this is actually the case. Computer simulations typically span only a few nanoseconds or less. If the simulation time is too short, the range of motion seen over the course of the simulation may be a poor representation of the actual flexibility of the native protein.

(4) The use of B-factors to infer atomic fluctuations is based on a very strong assumption: that the fluctuations are isotropic and Gaussian. If the true atomic motion is very different from this, then B-factors may give a poor estimate of its magnitude.[12]

In this article, we provide answers to these questions by studying the atomic fluctuations of the small protein bovine pancreatic trypsin inhibitor (BPTI). First, we perform simulations of the protein, both in solution and in crystal, in order to estimate thermal fluctuations under both sets of conditions. We then compare these estimates to the fluctuations derived from room temperature crystallography data by two different methods: conventional refinement using isotropic B-factors, and a recently developed constrained multiple copy refinement method.[13] We do not directly compare these results to NMR data, since the ambiguities in the interpretation of NMR makes a direct comparison difficult. However, we occasionally make reference to the results of previous NMR studies, when those results help to clarify the phenomena we observe.

A. Solution versus crystal

The first two issues mentioned above can be studied by simulating a protein in crystal form rather than in solution. Of course, the simulation of a complete protein crystal is out of the question. To model even a single protein molecule requires a considerable amount of computer time, and a crystal of any size would be far beyond the limits of any currently existing computer. On the other hand, we can create a reasonable imitation of a crystal by modeling a single crystallographic unit cell, then applying periodic boundary conditions.[14] This should give an accurate representation of the interactions between adjacent molecules.

This also allows us to characterize the global motion of each molecule within the crystal. It does not completely eliminate the need for global motion correction, since there is nothing to prevent the entire unit cell from drifting parallel to the walls of the simulation box. However, this correction can be applied to the unit cell as a whole, while retaining the motion of individual molecules relative to each other. There is no need to correct for global rotations, since the periodic boundary conditions enforce a set of crystal axes on the model.

B. Constrained multiple copy refinement

An alternative to using isotropic B-factors for refining crystallographic structures is a method known as multiconformer refinement.[15] Rather than fitting a single structure to the experimental data, it uses an ensemble of structures, which are simultaneously minimized so that their composite electron density provides a best fit to the experimental diffraction data. Atomic fluctuations can be derived from this method by calculating the variation in the position of each atom over the ensemble of structures. It is no longer necessary to make any assumptions about the distribution of the fluctuations. Therefore, whenever the motion differs significantly from being isotropic and Gaussian, multi-conformer refinement is expected to give a more realistic characterization than isotropic B-factors.

This method has one serious problem. The use of multiple structures to fit the experimental data greatly increases the number of fitting parameters. This means that it can only be used with very high quality data sets. Otherwise, the ratio of fitting parameters to data points becomes unreasonably large, and the method merely fits to the noise in the data.

A new method has recently been developed, known as constrained multiple copy refinement,[13] which solves this problem. It uses an ensemble of structures, usually about ten, to fit the experimental data (Fig. 1). To maintain an acceptable ratio of fitting parameters to data points, all bond lengths and angles are held fixed at their ideal values. In test cases, this method has been found to give a better fit than conventional, single copy refinement methods, while using roughly the same number of fitting parameters. In comparisons of the two methods on room temperature diffraction data, the atomic fluctuations obtained by multiple copy refinement are usually found to be larger than those obtained from isotropic B-factors, sometimes by as much as a factor of 5 for side chain atoms.

C. Bovine pancreatic trypsin inhibitor

BPTI is a 58 residue protein belonging to the Kunitz class of inhibitors. A number of crystal structures have been determined for it,[16,17] including several structures of BPTI complexed with other proteins.[18,19] The core of the protein

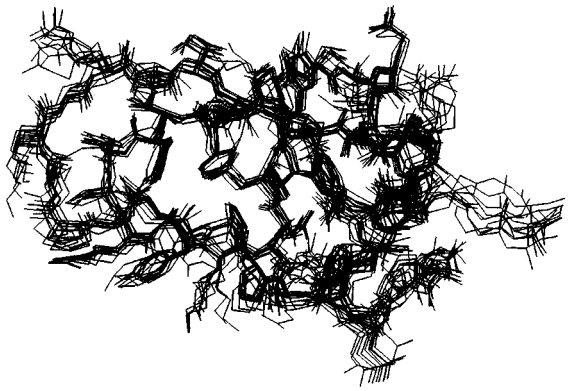

FIG. 1. Ten structures of BPTI, obtained by constrained multiple copy refinement of crystallography data (Ref. 13). The data was provided by Huber (see Ref. 17). See discussion in Methods section.

consists of a triple stranded β sheet. A pair of helices are found at the ends of the chain. It is a highly stable protein, containing three sulfur bridges.

Residues 11–19 compose the principle binding loop.[18] This loop attaches to the enzyme binding pocket, forming an intermolecular antiparallel β-sheet. A secondary binding loop, residues 34–39, also contributes to protease binding.

D. Multiple time step langevin dynamics

To perform the simulations described in this work, we use a multiple time step Langevin dynamics algorithm, which significantly reduces the amount of CPU time required.[20] The principle behind simulation algorithms of this type is that very fast motions, such as bond length and angle oscillations, are not of interest. Therefore, rather than trying to accurately calculate the motions of these very fast degrees of freedom, we can save time by simply treating them stochastically. They then act as a heat reservoir for the rest of the system.

This separation is accomplished by using time steps of various sizes. For the longest steps, we explicitly constrain all bond lengths and angles to remain fixed. This allows a much larger time step than would otherwise be possible. The highly constrained steps are interspersed with shorter, unconstrained steps which allow the fast degrees of freedom to remain in thermal equilibrium.

This may be contrasted with the multiple time step Verlet algorithms which have gained popularity in recent years.[21,22] These algorithms attempt to save CPU time by recalculating some forces less often than once every time step. Because they still integrate all degrees of freedom on the fastest time scale present, they can never reduce the number of time steps required compared to conventional molecular dynamics. At best, they can reduce the CPU time required for each time step by economizing on force evaluations. In contrast, the multiple time step Langevin dynamics method actually reduces the number of time steps required, by ignoring the dynamics of the fastest degrees of freedom. As a result, we estimate that our algorithm is 2–3 times faster than multiple time step MD algorithms.

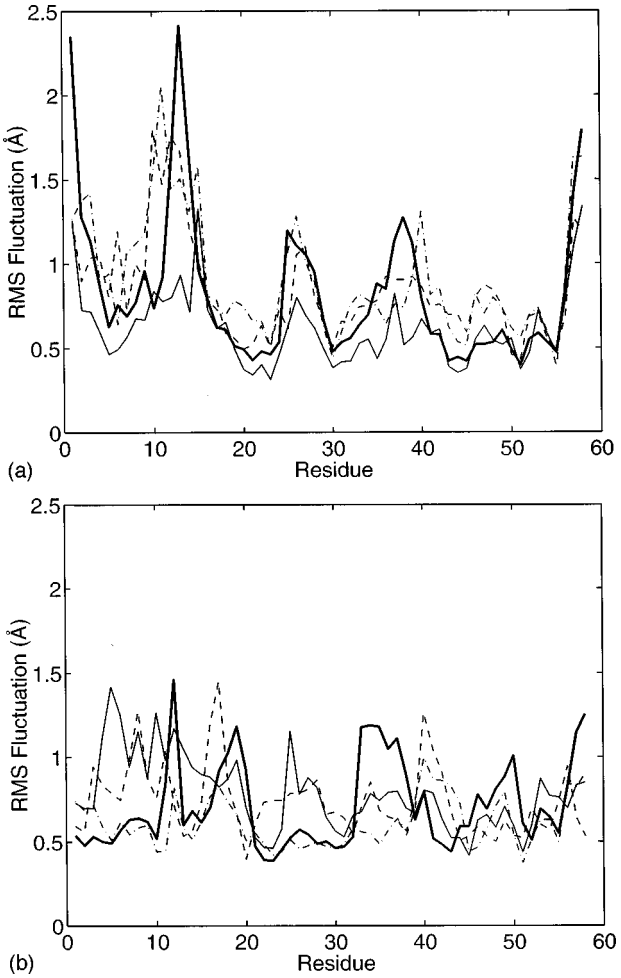

FIG. 2. Root mean square fluctuations of C_α atoms about their average positions. (a) Fluctuations for the four solution simulations. (b) Fluctuations for the four protein molecules in the crystal simulation. For the solution simulations, all global motion of each molecule has been removed. For the crystal simulation, only global translations of the entire unit cell have been removed.

II. RESULTS AND DISCUSSION

A. Convergence of fluctuations

We have performed five simulations of BPTI: one of a crystallographic unit cell containing four protein molecules, and four of a single protein molecule in solution. All of the simulations are approximately 4 ns in effective duration. The proteins remain quite stable over the course of the simulations. The final root mean square (rms) C_α deviations from the starting structure range from 1.64 to 2.49 Å for the four molecules in solution, and from 1.59 to 1.78 Å for the molecules in the crystal (after removing all global motion of each molecule). These deviations are similar to those found in other nanosecond scale simulations. For example, a 1 ns simulation of BPTI[23] found a deviation from the starting structure of 1.56 Å for the first 56 residues, while a 1.4 ns simulation[3] found deviations as large as 1.81 Å.

The rms fluctuations of the C_α atoms about their average positions are shown for the four solution simulations in Fig. 2(a), and for the four protein molecules of the unit cell in Fig. 2(b). For the solution simulations, global motion has

been removed by rotating and translating each structure to minimize the rms C_α deviation from the starting structure. For the unit cell simulation, only global translations of the whole unit cell have been removed.

Before we attempt to learn anything from this data, we must first determine whether the simulations were long enough for the results to have converged to their final equilibrium values. Otherwise, our data will not represent an adequate sampling of conformation space, and it will be difficult if not impossible to draw reliable conclusions from it.

For the solution simulations, the agreement between the four curves is reasonably good. There are some differences in the details, particularly in the less stable *N*-terminal region, but overall it is clear that the results have begun to converge. We can therefore conclude that we have obtained a reasonable sampling of those fluctuations which occur on a time scale of 4 ns or less. We must be careful not to draw too strong a conclusion, however, and assume that we have sampled the full range of motion available to the protein. The length of the simulations must always be kept in mind. If the real protein experiences conformational transitions on a time scale which is long compared to this, we would not expect those transitions to occur during the course of our simulations. We would then expect our simulations to converge toward each other, but the converged fluctuations would still not represent the full flexibility of the native protein.

The agreement between the four molecules of the unit cell is much poorer. All four curves show fluctuations of about the same magnitude, but the details of the curves differ considerably from one molecule to another. Once again, the differences are largest in the *N*-terminal region, roughly the first 20 residues. This reflects the nature of the protein: the binding loop is considerably more mobile than the central β sheet. Even in the core of the protein, however, we see significant differences between molecules.

There are two obvious explanations for this discrepancy. First, it is possible that crystal contacts inhibit conformational changes of each molecule. The equilibration process would then be slowed down relative to a molecule in solution, and the results of the crystal simulation would be less well converged than the solution simulations. Second, it is possible that the simulation breaks the symmetry between the four molecules. That is, they drift relative to each other so that they are no longer exactly symmetrically positioned within the unit cell. Each molecule then sees slightly different surroundings, and hence displays slightly different fluctuations.

There is some evidence for the latter explanation. If symmetry were being exactly preserved, we should be able to treat all four molecules as equivalent (after applying the appropriate symmetry transformation to superimpose them). We could then calculate a single set of fluctuations from the complete ensemble of conformations of all molecules. In fact, if we do this the fluctuations increase by roughly an angstrom compared to those shown in Fig. 2(b). This indicates that the molecules have moved such that their positions are no longer exactly symmetric.

Presumably this is an artifact, either of the potential function or, more likely, of insufficient equilibration. The true thermodynamic ensemble should exhibit the symmetry which is observed experimentally in the crystal. If we could run the simulation long enough, the molecules should return to symmetrical positions, and the local fluctuations of the four molecules should become identical. However, this might take quite a long time. At present, the best we can do is to average the four curves and take this as a best estimate of the fluctuations. We do this for both the solution and crystal simulations, and in the rest of this article, it is these average curves which appear in all figures. The differences between the curves shown in Figs. 2(a) and 2(b) can then be taken as an estimate of the uncertainties in our results.

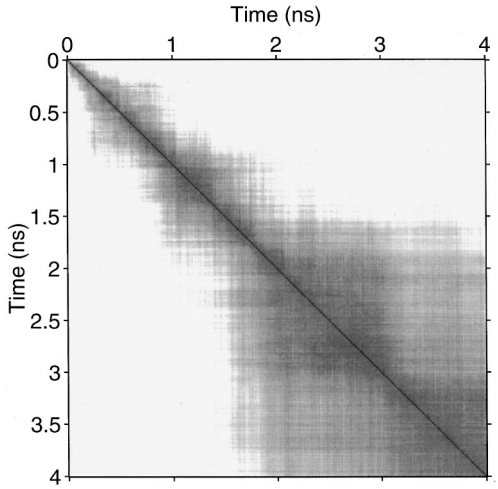

FIG. 3. Deviation matrix for the crystal simulation. Black indicates identical structures, while structures which differ by more than 1 Å are indicated by white. Intermediate shades of gray are used for deviations between 0 and 1 Å.

B. Deviation matrix analysis

Another method for assessing the exploration of conformation space is to calculate the rms C_α deviation of every structure which occurs during the course of the simulation from every other. The resulting matrix of deviations can then be plotted as a two-dimensional image, as in Fig. 3. Black corresponds to identical structures, while structures which differ by more than 1 Å are indicated by white. Metastable states where the system remains for extended periods show up as dark squares along the diagonal. Dark patches away from the diagonal indicate that the system has moved away from a configuration, then returned to it again.

Troyer and Cohen[24] used this method to study a 1.02 ns simulation of BPTI. They found that the protein moved through a series transient states, undergoing small fluctuations within each state before moving on to the next one. In 1 ns, they never saw it return to a state which it had visited before.

In our crystal simulation, we see similar behavior. It is also clear, however, that as the simulation progresses, the system spends more time in each state before moving on to the next one. The entire second half of the simulation is spent in only two major states. The result is that, if we cal-

culate atomic fluctuations based only on the second half of the simulation, we obtain values significantly smaller than those shown in Fig. 2(b).

This is, essentially, the time scale problem mentioned in the introduction. Can a simulation spanning only a few nanoseconds explore the same range of motion seen in an experiment which lasts a billion times longer? Furthermore, if the motion varies significantly from one part of the simulation to another, then which part is a better representation of the thermodynamic ensemble?

Based on our results, it appears that a simulation can be divided into two phases. Initially, the system is not trapped in any sort of local energy minimum. It therefore moves freely through a variety of transient states, with each part of the protein exploring a wide range of motion. Eventually, however, it settles down into a metastable local minimum, and remains there for an extended period. If we are still to sample the protein's full range of motion, we need a simulation which is long compared to the lifetimes of all local energy minima. It is well known from NMR relaxation measurements that proteins exhibit conformational transitions on the time scale of a few nanoseconds,[25,26] but this is only a lowest estimate. Indeed, there has been some experimental evidence for protein domain motions occurring on a time scale of roughly 1 microsecond.[26] Thus, a simulation lasting microseconds or longer might be required to fully explore the protein's range of motion.

A similar observation was made by Caves *et al.*[27] based on simulations of crambin in vacuum. They found that ten independent simulations, each spanning 120 ps, collectively explored a larger region of conformation space than a single simulation spanning 5 ns. Consequently, they recommended using multiple short simulations, rather than a single long one, for conformational sampling.

Although our results tend to support this strategy, they also show that it must be applied with care. The early, unstable part of a simulation exhibits behavior which is qualitatively different from the later parts. Furthermore, it is heavily influenced by the starting conditions, and may visit regions of conformation space which are not significantly populated under equilibrium conditions. Several short simulations may explore a larger region than a single long one; but there is no *a priori* guarantee that they offer a better representation of the true thermodynamic ensemble.

This leads to an important question. How similar are atomic fluctuations during the two phases? Clearly the rates of motion are very different. Initially, the system moves rapidly through a variety of transient states. Eventually it settles down into longer lived, metastable states, and experiences only occasional transitions between them. But are the magnitudes the same? Does the protein experience the same overall range of motion in both phases?

At present, the only answer we can give is to compare the simulated fluctuations to experiment. It will be seen that in many cases, they agree reasonably well. We therefore offer a tentative ''yes'' to the above questions, at least for the most stable parts of the protein. Ultimately, the only way to obtain a definite answer will be to perform much longer simulations, hundreds of nanoseconds or more, and find out.

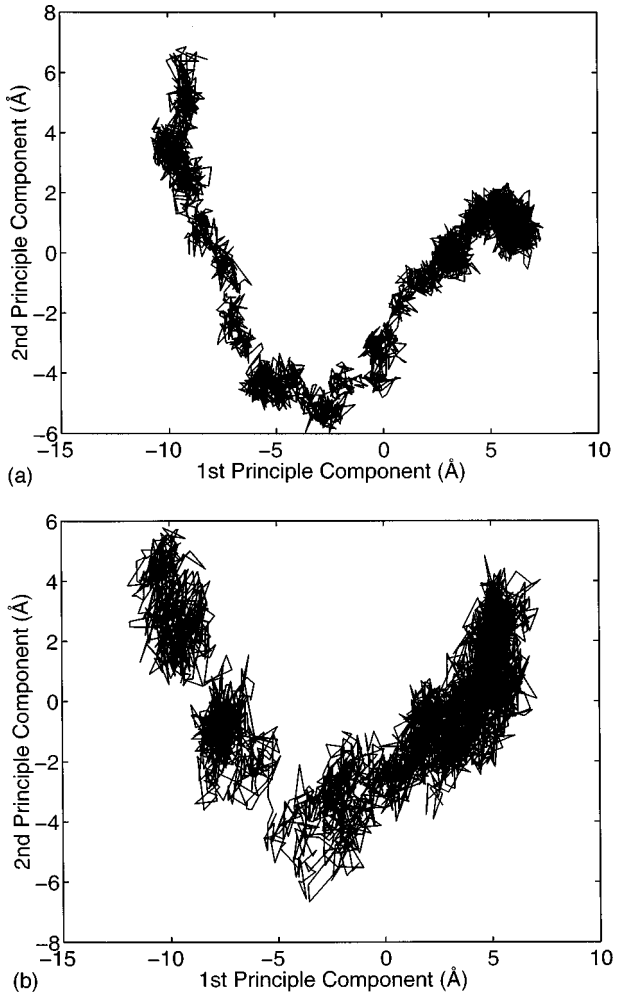

FIG. 4. Simulation trajectories, plotted in the plane of the first two principle components. (a) The crystal simulation. (b) One of the four solution simulations.

C. Principle component analysis

Still another method for studying the behavior of a simulation is principle component analysis.[27–29] This involves diagonalizing the variance-covariance matrix for atomic displacements calculated over a given simulation. The eigenvectors are known as principle components, and represent coordinated motions of the entire system. The corresponding eigenvalues indicate how much each principle component contributes to the total mean square fluctuations of the entire system. It often is found that only the first few principle components account for the majority of the atomic fluctuations.[29] These components then define an ''essential subspace,'' in which most of the important motions occur.

We have performed a principle component analysis of each of the five simulations. Figure 4(a) shows the trajectory of the crystal simulation, plotted in the plane defined by the first two principle components. Figure 4(b) is a similar plot for one of the solution simulations. (Only one solution simulation is displayed, since all four exhibit similar behavior.) The first two principle components account for 52% of the total mean-square fluctuations in the case of the crystal simulation, and 64% in the case of the solution simulation.

The transient and metastable states discussed above now

show up as a ''chain of beads.'' The spatial extent of each state is clearly visible, as are the transition paths linking them together.

The difference between the two plots is quite striking. They both show the chain of beads pattern, but in the crystal simulation, the beads are much smaller, and there are many more of them. It appears that crystal contacts significantly alter the energy landscape in which the protein moves. Metastable states may be thought of as valleys in the energy landscape. The crystal contacts create barriers within the valleys, forcing the protein to find its way through a rougher landscape characterized by a larger number of local minima. This may account for the slower equilibration discussed above.

One might reasonably wonder whether the different behavior observed for the crystal simulation could simply reflect the greater number of molecules present. After all, a larger system would be expected to have more local minima, even if the energy landscape experienced by each molecule were no different than in solution. This can easily be tested by performing the principle component analysis on only a single protein molecule within the unit cell. The results remain quite different from the solution simulations (data not shown). The local minima remain much smaller than in solution, and although there are fewer of them than in Fig. 4(a), there are still more than in solution. This indicates that crystal contacts really do change the character of the energy landscape.

Indeed, the first principle component for the complete unit cell shows significant displacements for atoms in all four protein molecules. This indicates that motions of the four molecules are highly correlated. The dominant motions in the crystal appear to be collective fluctuations of many protein molecules, which are likely to be quite different from those in solution.

D. Solution versus crystal

We now examine the differences between the atomic fluctuations of a protein in solution and in a crystal. This issue has important consequences for the interpretation of crystallographic data, and for the comparison of crystallography to other methods, such as NMR and computer simulations, which deal with proteins in solution.

In Fig. 5, we plot the mean C_α fluctuations for solution (thick line) and crystal (dashed line). Over much of the protein, we see that the two actually agree quite well with each other. The differences are limited primarily to four regions: residues 1–3, 9–15, 25–28, and 57–58. All of these are loop regions which are relatively disordered in the solution simulations. The crystal contacts reduce the fluctuations in these unstable regions, while having little effect on the rest of the protein.

As discussed earlier, the fluctuations calculated from solution and crystal differ in two respects. First, crystal contacts inhibit the internal motion of each molecule. Second, a molecule in a crystal experiences a small amount of global motion (translations of and rotations about the center of mass). This motion is present in experimental data, and cannot easily be separated from the more interesting internal motions.

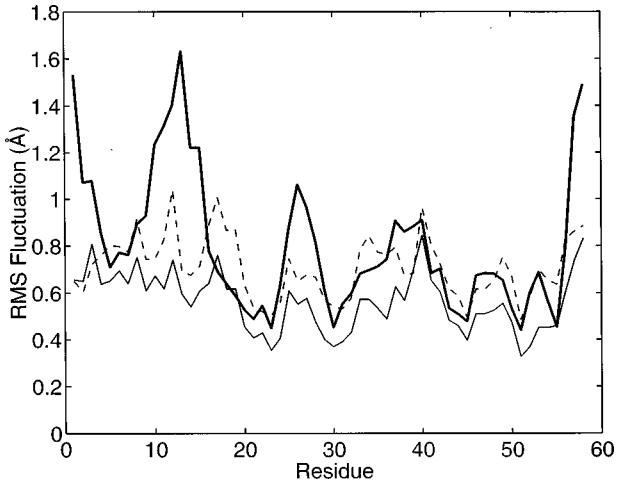

FIG. 5. Mean C_α fluctuations calculated from the solution simulations (thick line), crystal simulation (dashed line), and crystal simulation with the global motion of each molecule removed (thin line). The difference between the thin and dashed lines indicates the magnitude of global motion within the crystal. The difference between the thick and thin lines indicates the effect of crystal contacts.

We can distinguish these two effects by removing all global motion present in our crystal simulation. We do this by rotating and translating each molecule to minimize its rms C_α deviation from its starting conformation. We then calculate rms fluctuations for each molecule, and as before, average the four curves. The result is shown as the thin line in Fig. 5.

The difference between the thin and dashed lines indicates the magnitude of global motion within the crystal. It has an rms of 0.17 Å, and is relatively uniform throughout the protein. For no residue is it more than 0.32 Å. To a first approximation, global motion seems to do little more than add a constant to the fluctuations everywhere. Of course, the size of this contribution might change if we increased the length of the simulation. NMR measurements on BPTI in solution have yielded global rotational correlation times in the range of 2–8 ns.[23] The numbers will undoubtedly be different for motion in crystal; however, if we take this as a rough estimate of the time scale on which global motions occur, it seems likely that 4 ns is not sufficient to fully explore the range of global motion accessible to the protein.

The difference between the thick and thin lines indicates the effect of crystal contacts. This is a much larger contribution than global motion, with an rms of 0.35 Å. It also varies considerably between residues, being essentially zero for some residues, and as large as 1 Å for others. Once again, we note that crystal contacts have the largest effect on the highly mobile loop regions, whose motion is significantly reduced in the crystal, as shown in Fig. 6.

It is especially worth noting that part of the principle binding loop has significantly reduced flexibility in crystal compared to solution. It is not surprising that this region should be highly flexible, given its ability to bind to a variety of substrates. Indeed, in the absence of any other information about protein function, this loop's flexibility in solution would be an important clue to its functional role. On the other hand, if one were relying on crystallography to deter-

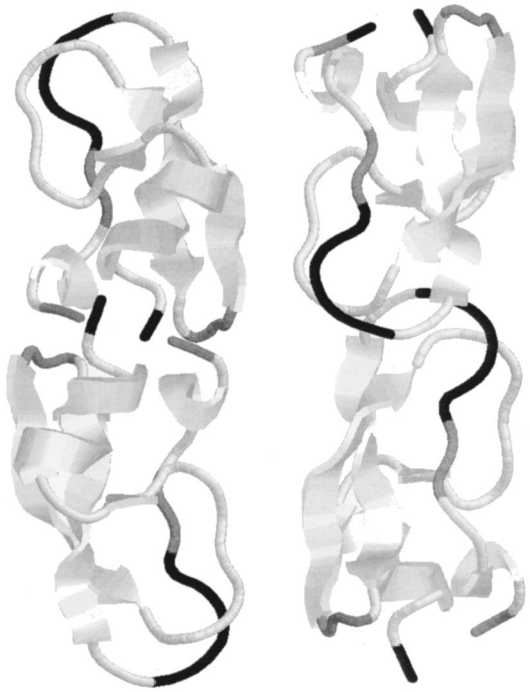

FIG. 6. A unit cell of BPTI. Residues are shaded according to how much their fluctuations are reduced by crystal contacts (the difference between the thick and thin lines in Fig. 5). Differences of more than 0.6 Å are indicated by black. Gray is used to indicate differences of between 0.3 and 0.6 Å, and white for differences of less than 0.3 Å.

mine flexibility, this clue would be completely missed. Neither the simulated nor experimental data show any unusual amount of motion for this loop in crystal.

We also note that the long duration of the simulations is critical for observing this effect. Earlier studies which tried to calculate the effect of crystal contacts based on much shorter simulations found little difference between protein flexibility in solution and in crystal.[8]

Another interesting difference between the protein's behavior in the solution and crystal simulations becomes apparent when one calculates the radius of gyration, as shown in Fig. 7. In solution, r_g decreases by approximately 0.2 Å, while in the crystal it increases by about the same amount. These changes are small, but they are consistent enough from one molecule to another to indicate that this is a real effect.

The fact that r_g expands in the crystal presumably represents inaccuracies in the model or the potential function; the starting structure is derived from a crystal, and ideally the simulation should reproduce it. Small deviations like this are to be expected, given the approximate nature of the computer model. On the other hand, the difference in behavior between the crystal and solution simulations is significant. The same potential function is used in both cases, so any differences reflect actual effects of the protein environment.

One possibility is that the change in r_g merely reflects a slight pressure difference between the two models. However, we would expect the pressure to be higher in the crystal, due to the smaller available volume per molecule, and hence for the radius of gyration to be smaller. In fact, we see the exact opposite. On the other hand, because we use only a limited

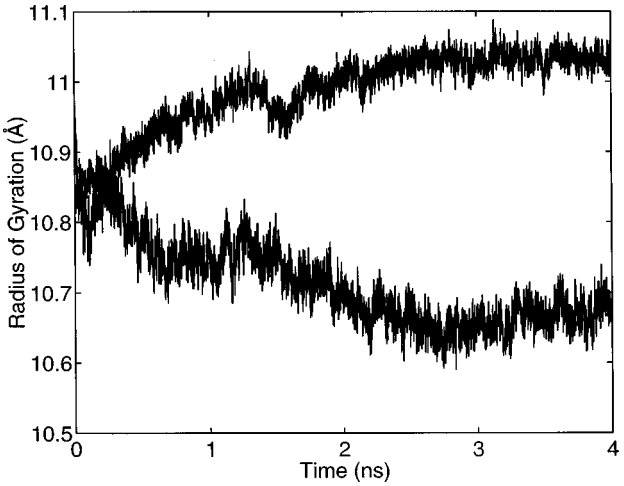

FIG. 7. Radius of gyration as a function of time, averaged over the four molecules. The data is from the crystal simulation (upper curve) and the solution simulations (lower curve). Both sets of data can be reasonably fit by exponentials with time constants of 1.14 ns (crystal) and 1.34 ns (solution).

shell of waters to hydrate the protein molecules, it is conceivable that surface tension in the solution simulations causes the pressure to actually be higher than in the crystal simulation.

Another possible explanation is that this is due to the hydrophobic effect. The protein contracts slightly in solution, so as to minimize its nonpolar surface area. Conversely, the protein in crystal expands slightly, due to a hydrophobic attraction between adjacent molecules.

E. B-factors versus multiple copy refinement

Figure 8 shows the rms C_α fluctuations calculated from the crystal simulation, along with those obtained experimentally by both methods of analysis, isotropic B-factors and constrained multiple copy refinement. It is apparent from visual inspection that all three methods have a reasonable

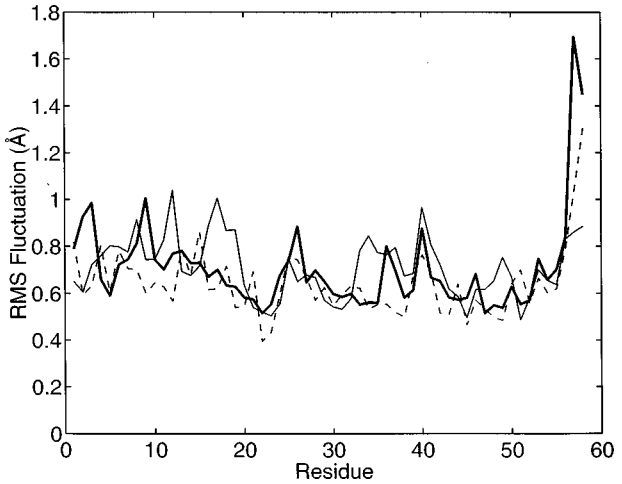

FIG. 8. C_α fluctuations calculated from the crystal simulation (thin line), and from analysis of crystallographic data by single copy refinement with isotropic B-factors (dashed line) and constrained multiple copy refinement (thick line).

TABLE I. The root mean square deviations between the atomic fluctuations calculated from the crystal simulation, single copy refinement with isotropic B-factors, and constrained multiple copy refinement (CMCR). The columns show the results when the calculation is based on all α-carbons, only the α-carbons in the core of the protein (residues 20–56), all heavy atoms, and heavy atoms in the core of the protein. These numbers give the deviations between the various curves shown in Figs. 8 and 9, not to be confused with the rmsd between pairs of protein structures.

	all C_α	core C_α	Heavy atoms	Core heavy atoms
B-factor vs simulation	0.17	0.14	0.24	0.21
CMCR vs simulation	0.19	0.11	0.28	0.26
B-factor vs CMCR	0.15	0.09	0.17	0.16

agreement with each other. They all show fluctuations of about the same magnitude, and there are significant correlations between the different curves.

We can make these observations more quantitative by calculating the root mean square deviations betweeen the various curves. These are shown in Table I. Overall, the B-factors give a better fit to the simulation data than does multiple copy refinement. On the other hand, the agreement between the two methods of crystallographic data analysis is significantly better than the agreement of either one with the simulation data. This suggests that uncertainties in the simulation data are the single largest source of error. It is probably more meaningful, therefore, if we restrict our analysis to the core of the protein, residues 20–56, in which the simulation seems to be more fully converged. In this region, the agreement between all of the curves is significantly improved. Furthermore, the simulation data now comes closer to multiple copy refinement than to the B-factors, with a deviation of only 0.11 Å.

This still does not tell us what we really want to know. We would like to answer the question, how much of an error do we make by using isotropic B-factors as a measure of atomic motion? In the case of the α carbons, the answer appears to be, very little. This is not surprising. Motion of the backbone tends to be highly constrained. Modeling this motion with an isotropic Gaussian is probably a reasonable approximation.

For flexible side chains, the answer is very different. Figures 9(a)–9(c) show the fluctuations for all heavy atoms, derived from the experimental data by both methods of analysis, along with the corresponding values from the crystal simulation. In Fig. 9(a), we see that for the most part, the two refinement methods agree with each other quite well, rarely differing by more than 0.2 Å. On the other hand, when they do differ significantly, it is invariably the multiple copy refinement which gives larger fluctuations. In some cases they are dramatically larger, all the way to 3.56 Å for the N_ζ of Lys46. There are a total of 25 atoms with fluctuations larger than 1.38 Å, the largest value obtained for any atom using isotropic B-factors. For these highly mobile atoms, the assumption of isotropic Gaussian fluctuations breaks down, leading the B-factors to underestimate the degree of atomic motion.

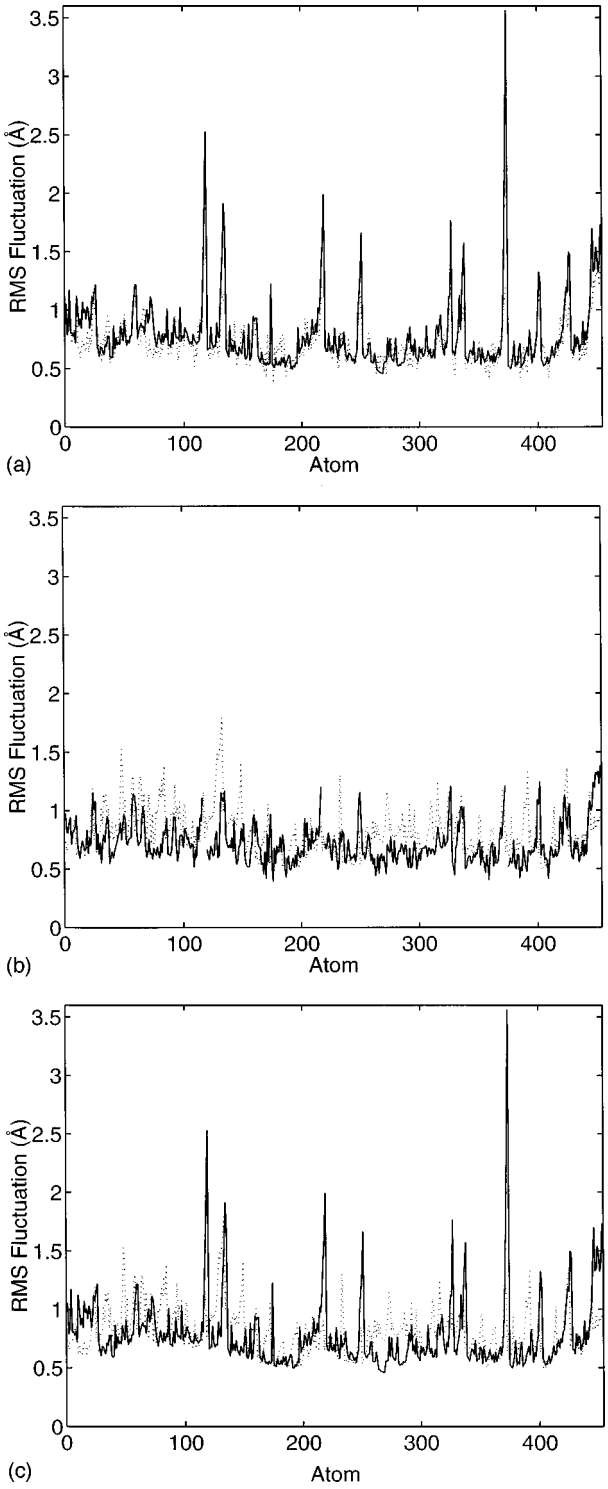

FIG. 9. Atomic fluctuations for all heavy atoms, calculated from the crystal simulation, B-factors, and multiple copy refinement. (a) Multiple copy refinement (solid line) and B-factors (dotted line). (b) B-factors (solid line) and simulation (dotted line). (c) Multiple copy refinement (solid line) and simulation (dotted line).

Interestingly, very few of these large fluctuations show up in the simulation data. In fact, when we consider all heavy atoms, the simulation gives a significantly better fit to

B-factors than to the multiple copy refinement, even if we restrict ourselves to the core residues.[30]

It is possible that this reflects a previously undiscovered problem with the multiple copy refinement method. A much more likely explanation, however, is that we are simply running into the time scale problem again. We previously posed the question of whether a simulation, spanning only a few nanoseconds, can explore the full range of motion of the native protein. For the most stable regions, the answer appears to be yes. It is hardly surprising if the answer is different for the flexible regions. The evidence suggests that 4 ns simply is not enough time for the long, highly exposed sidechains to fully explore their available range of motion. Consequently, the simulation sees only small fluctuations for these atoms, ones which are similar in magnitude to those obtained from isotropic B-factors.

It is also possible that some of the discrepancy between the simulation and crystallography results is due to the presence of static disorder in the crystal. It is likely that some of the disorder observed in crystallographic data comes from crystalline defects which do not change with time. A simulation which incorporates only a single unit cell obviously does not include this static disorder, so it is reasonable to expect some of the fluctuations to be smaller.

III. CONCLUSIONS

We have performed 4 ns simulations of BPTI both in crystal form and in solution. We have calculated atomic fluctuations from both sets of simulations, as well as those obtained experimentally by x-ray crystallography at room temperature using two different analysis methods: single copy refinement with isotropic B-factors, and constrained multiple copy refinement. By comparing these data sets, we have attempted to answer several questions about the nature of atomic fluctuations, and the various methods used to study them.

We find that crystal contacts have a significant effect on atomic motion. Fluctuations tend to be significantly larger in solution than in the crystal, sometimes by as much as 1 Å. The effect is most pronounced in the flexible loop regions, which are relatively disordered in the solution simulations. In contrast, the stable core regions of the protein are only slightly affected by crystal contacts.

The result is that the spectrum of fluctuations is considerably flatter in the crystal, without the large variations that are seen in solution. Attempting to determine a protein's inherent flexibility from its fluctuations in a crystal may therefore be deceptive. This is potentially a serious problem, since variations in flexibility are an important piece of information for relating structure to function.

Global motion correction is a less important factor. Global motion within the crystal increases the atomic fluctuations for all residues by roughly 0.1–0.2 Å, with little variation between different regions of the protein.

The use of B-factors to infer atomic fluctuations is a reasonable approximation for backbone atoms, especially in the most stable regions of the protein. For flexible sidechains, however, it can significantly underestimate the degree of motion, sometimes by several angstroms. We believe that constrained multiple copy refinement provides a more realistic assessment of the fluctuations of these highly mobile parts of the protein.

Finally, the gap in time scale between simulation and experiment must always be kept in mind. Early in a simulation, the protein is quite mobile and moves rapidly through a variety of transient states. One should not assume, however, that this is typical behavior. It may be only an artifact of the starting conformation. Within a few nanoseconds, it begins to settle into metastable states and remain in them for extended periods.

There is no guarantee that fluctuations calculated from the early, unstable phase of the simulation will agree with the true equilibrium ones. This is a question which can only be settled by performing much longer simulations. Based on the comparison with experimental data, it appears that the agreement is reasonably good for the most stable parts of the protein. For the less stable regions, however, a short simulation is able to explore only a small part of the full range of motion accessible to the protein.

IV. METHODS

We have performed five simulations of bovine pancreatic trypsin inhibitor: one of a crystallographic unit cell containing four protein molecules, and four of a single protein molecule in solution. The Amber/OPLS potential function[31,32] was used. A small van der Waals term ($\sigma = 1.0$ Å, $\epsilon = 0.01$ kcal/mol) was added to all hydrogen atoms to avoid a catastrophic divergence in energy whenever a hydrogen comes too close to any other charged atom. The temperature was set at 292 K.

Noncovalent interactions were cut off at 11 Å by means of a shifting function.[33] For future work, we plan replace the cutoff with a more accurate method for calculating long range interactions, such as the particle mesh Ewald method.[14] This should help the system to stay closer to the crystal structure, although the precise amount of improvement reported with this method has varied widely.[14,34] We do not expect it to qualitatively change any of our conclusions, however.

As a starting point, we used the 1BPI crystal structure from the Protein Data Bank,[16] including the 167 crystallographic water molecules. Six chloride ions were added to make the system electrically neutral. This number of water molecules corresponds to a water to protein mass ratio of 0.46. It is well known experimentally that most proteins are fully hydrated at a mass ratio of approximately 0.38.[35] Adding water beyond this point merely dilutes the protein without changing it in any way. Computer simulations by Steinbach and Brooks[36] showed similar results: adding water beyond this point did not cause any change in a variety of quantities, such as the deviation from the crystal structure and the rate of dihedral transitions. We therefore expect the system to be well hydrated. The use of a limited hydration shell, such as this one, has recently gained popularity as a computationally efficient way to model proteins in solution.[37–39]

For the solution simulations, a weak constraining potential was added to prevent water molecules from evaporating:

$$F_{\text{con}}=0 \quad (r<20\ \text{Å})$$
$$=-k(r-20\ \text{Å}) \quad (r>20\ \text{Å}), \tag{1}$$

where r is the distance of each atom from the origin, and where the force constant k was set at 0.1 kcal/mol Å^2.

For the crystal simulation, three additional copies of the system were created according to the crystallographic symmetry group, yielding a total of four protein molecules, 668 water molecules, and 24 ions. Periodic boundary conditions were applied with the unit cell dimensions given in the PDB file, and the entire system was energy minimized.

The multiple time step Langevin dynamics algorithm is built on the Langevin equation:

$$\mathbf{M}\ddot{\mathbf{X}}=-\eta\mathbf{M}\dot{\mathbf{X}}+\mathbf{F}(\mathbf{X})+\mathbf{B}, \tag{2}$$

where $\mathbf{X}$ is a $3N$ dimensional vector containing the coordinates of the atoms, $\mathbf{M}$ is the diagonal mass matrix, η is a friction coefficient, $\mathbf{F}=-\nabla E$, where E is the potential energy of the system, and $\mathbf{B}$ is an uncorrelated random force with zero mean and magnitude given by

$$\langle B_i B_j\rangle=2M_i k_B T\eta\delta_{ij}/\Delta t, \tag{3}$$

where k_B is Boltzmann's constant, T is the temperature, and Δt is the time step used in the simulation. The friction constant η may be viewed as a parameter which determines how strongly the system is coupled to an external heat reservoir.

If the system is very heavily damped, the term involving $\ddot{\mathbf{X}}$ may be omitted altogether. This yields the diffusive, or overdamped, form of the Langevin equation:

$$\eta\mathbf{M}\dot{\mathbf{X}}=\mathbf{F}(\mathbf{X})+\mathbf{B}. \tag{4}$$

We can eliminate η from this equation by rescaling the units of time, so that $\eta=1$. Time is then measured in units of a characteristic time τ_0. In the limit of sufficiently small timestep, both Eqs. (2) and (4) can be shown to produce a canonical probability distribution.[40,41]

In our earlier work, we used Eq. (4) for both the long, highly constrained steps, and the shorter, unconstrained ones. In principle, this yields a savings in CPU time of nearly a factor of 100. In practice, however, we find that this is an inefficient way of simulating a protein. The problem is this: real proteins are not highly damped on the time scale of individual MD steps. Thus, by explicitly overdamping every mode of the protein, we severely impair its ability to explore conformation space.

This can be understood by considering the simple case of a random walk in one dimension. If every step has length L, and is uncorrelated with the previous step, then the mean distance traveled in N steps is $L\sqrt{N}$. Suppose, however, that the steps are correlated: that the system always takes M steps in a row in the same direction, before randomly chosing a new direction. In this case, the mean distance traveled in N steps is $L\sqrt{MN}$. For large M, this leads to a much more rapid exploration of configuration space.

In the case of a protein, there are modes which remain correlated over hundreds of MD steps. Furthermore, these modes are generally the ones which contribute most to large scale conformational changes. Eliminating their correlations from one step to the next will dramatically slow down the exploration of configuration space.

In our current work, we solve this problem by using the full Langevin Eq. (2) for the long, highly constrained steps. We still use Eq. (4) for the shorter steps, since bond length and angle oscillations have short correlation times, and contribute little to the large scale conformational changes.

Each iteration of the multiple time step Langevin dynamics algorithm consists of three time steps:

(1) A ''long'' step, using the full Langevin Eq. (2), with all bond lengths and angles constrained, and a time step of 10 fs.
(2) A ''medium'' step, using the diffusive Langevin Eq. (4), with bond lengths but not angles constrained, and a time step of $2.5\times10^{-4}\tau_0$.
(3) A ''short'' step using Eq. (4), with no constraints, and a time step of $5\times10^{-5}\tau_0$.

The long steps are calculated using a leapfrog style Verlet integrator. The short and medium steps are calculated with a simple first order Euler integrator.[33]

The CPU time required is further reduced by dividing the nonbonded interactions into ''strong'' and ''weak'' components. The strong part, consisting of interactions between atoms closer than 4.5 Å, is recalculated for every time step, while the remaining interactions are recalculated only once per iteration.

All of the simulations were 4×10^5 iterations in length, or 4 ns if we assume that each iteration corresponds to 10 fs. (The true figure is slightly longer than this, since some motion of the slow degrees of freedom also occurs in the short and medium steps.) Structures were saved after every 100 iterations. For purposes of analysis, we treat the first 100 ps of each simulation as an equilibration period, and calculate all averages based on the remaining 3.9 ns.

It is not immediately obvious what value should be used for the damping constant in the long steps. As noted above, damping the system too heavily reduces its ability to explore conformation space. The slowest underdamped modes in proteins are known to have a period of approximately 1 ps,[42] so the damping constant must be small enough to avoid excessive damping on this time scale. Furthermore, the system is already weakly coupled to the Langevin heat reservoir by the short and medium steps. In these simulations, we set $\eta=1$ ps^{-1}, but a different value might ultimately be found to give better results.

The LINCS algorithm[43] was used for implementing the constraints. LU decomposition of the constraint matrices was done with the Sparse 1.3a package.[44]

Some difficulties were encountered constraining bond lengths and angles together. The problem is that, wherever four atoms are constrained to lie in a plane, the constraint matrix is ill conditioned. This results in very large constraint forces which are delicately balanced to cancel each other out. The problem is helped, but not eliminated, by attaching a dummy atom to each planar group.[45]

For the current work, we used the following inelegant but effective solution. We monitored the constraint forces in

each long step. If any one of them was larger than a cutoff value (set at 10^4 kcal/mol Å), we canceled that step, and performed an extra pair of short and medium steps to let the planar atoms move slightly out of alignment. For our future work, we hope to find a better method of implementing the constraints which avoids this problem altogether.

The presence of explicit water molecules creates some special difficulties which must be dealt with. Water librational motion has a period of approximately 42 fs.[22] To integrate this accurately, we therefore need a maximum time step of no more than about 5 fs.

This problem can be addressed by increasing the hydrogen mass used in the simulation, so as to slow down the water librations. A time step as large as 10 fs then becomes possible, allowing much more efficient exploration of configuration space. We therefore used a hydrogen mass of 5 amu for all of the simulations.

One might reasonably object that this is an unrealistic change, which will lead to incorrect molecular dynamics. The effect of such a change was studied by Pomes and McCammon,[46] who simulated a box of water with a hydrogen mass of 10 amu. They found that free energies and radial distribution functions were completely unaffected, but translational diffusion rates were reduced by 28%. We expect the corresponding rate changes for the water molecules in our simulations to be smaller than this, since we use a hydrogen mass only half as large. Any effect on the protein should be smaller still, since explicit hydrogens constitute only 25% of the protein atoms in the Amber/OPLS representation, compared to 2/3 of the water atoms. In any case, the quantities of interest in this work, the magnitudes of equilibrium thermal fluctuations, do not depend on atomic masses.

Principle component analysis was performed by diagonalizing the variance-covariance matrix for atomic displacements of all α-carbons, calculated from the last 3.9 ns of each simulation.[28] Global motion of each molecule was first removed by applying rotations and translations to minimize its rms C_α displacement from the starting structure. Thus, the principle components reflect only the internal motion within each molecule.

We obtained the room temperature crystallographic diffraction data which was used to generate the 4PTI crystal structure.[17] We then reanalyzed it using the constrained multiple copy refinement method with an ensemble of ten structures,[13] shown in Fig. 1. Only a single copy of the water molecules from the 4PTI file was used in the multiple copy refinement, and their coordinates were not further refined. For comparison, we also calculated atomic fluctuations from the B-factors given in the 4PTI pdb file according to[3]

$$\langle(\Delta r)^2\rangle^{1/2}=(3B/8\pi^2)^{1/2}. \tag{5}$$

We chose this data set, rather than the data set used to generate the 1BPI crystal structure, since the latter was collected at 125 K, in contrast to the simulations which were performed at room temperature.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Robert Huber for providing room temperature diffraction data for BPTI. S. Doniach wishes to thank H. Berendsen for a useful discussion and for pointing out the work by Hess et al. This work was supported by National Science Foundation Grant No. PHY 9418964. P. Eastman is a recipient of a National Science Foundation Graduate Research Fellowship.

[1] M. Gerstein, A. M. Lesk, and C. Chothia, Biochemistry **33**, 6739 (1994).
[2] R. Huber and W. S. Bennett, Biopolymers **22**, 261 (1983).
[3] P. H. Hünenberger, A. E. Mark, and W. F. van Gunsteren, J. Mol. Biol. **252**, 492 (1995).
[4] K. N. Trueblood, in *Accurate Molecular Structures*, edited by A. Domenicano and I. Hargittai (Oxford University Press, Oxford, 1992), pp. 199–219.
[5] H. Frauenfelder, G. A. Petsko, and D. Tsernoglou, Nature (London) **280**, 558 (1979).
[6] G. Lipari and A. Szabo, J. Am. Chem. Soc. **104**, 4546 (1982).
[7] C. L. Brooks, M. Karplus, and B. M. Pettitt, *Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics* (Wiley, New York, 1988).
[8] W. F. van Gunsteren and M. Karplus, Biochemistry **21**, 2259 (1982).
[9] J. Kuriyan and W. Weis, Proc. Natl. Acad. Sci. USA **88**, 2773 (1991).
[10] D. L. D. Caspar, J. Clarage, D. M. Salunke, and M. Clarage, Nature (London) **332**, 659 (1988).
[11] J. Doucet and J. P. Benoit, Nature (London) **325**, 643 (1987).
[12] A. E. Garcia, J. A. Krumhansl, and H. Frauenfelder, Proteins: Struct., Funct., Genet. **29**, 153 (1997).
[13] M. Pellegrini, N. Gronbech-Jensen, J. A. Kelly, G. M. U. Pfluegl, and T. O. Yeates, Proteins: Struct., Funct., Genet. **29**, 426 (1997).
[14] D. M. York, A. Wlodawer, L. G. Pedersen, and T. A. Darden, Proc. Natl. Acad. Sci. USA **91**, 8715 (1994).
[15] F. T. Burling and A. T. Brunger, Isr. J. Chem. **34**, 165 (1994).
[16] S. Parkin, B. Rupp, and H. Hope, Acta Crystallogr., Sect. D: Biol. Crystallogr. **52**, 18 (1996).
[17] A. Wlodawer, J. Deisenhofer, and R. Huber, J. Mol. Biol. **193**, 145 (1987).
[18] A. J. Scheidig, T. R. Hynes, L. A. Pelletier, and J. A. Wells, Protein Sci. **6**, 1806 (1997).
[19] J. J. Perona, C. A. Tsu, C. S. Craik, and R. J. Fletterick, J. Mol. Biol. **230**, 919 (1993).
[20] P. Eastman and S. Doniach, Proteins: Struct., Funct., Genet. **30**, 215 (1998).
[21] M. Tuckerman, B. J. Berne, and G. J. Martyna, J. Chem. Phys. **97**, 1990 (1992).
[22] T. Schlick, E. Barth, and M. Mandziuk, Annu. Rev. Biophys. Biomol. Struct. **26**, 181 (1997).
[23] P. E. Smith, R. C. van Schaik, T. Szyperski, K. Wüthrich, and W. F. van Gunsteren, J. Mol. Biol. **246**, 356 (1995).
[24] J. M. Troyer and F. E. Cohen, Proteins: Struct., Funct., Genet. **23**, 97 (1995).
[25] G. M. Clore, A. Szabo, A. Bax, L. E. Kay, P. C. Driscoll, and A. M. Gronenborn, J. Am. Chem. Soc. **112**, 4989 (1990).
[26] R. T. Clubb, J. G. Omichinski, K. Sakaguchi, E. Appella, A. M. Gronenborn, and G. M. Clore, Protein Sci. **4**, 855 (1995).
[27] L. S. D. Caves, J. D. Evanseck, and M. Karplus, Protein Sci. **7**, 649 (1998).
[28] S. Hayward, Kitao A, F. Hirata, and N. Gō, J. Mol. Biol. **234**, 1207 (1993).
[29] A. Amadei and A. B. M. Linssen, and H. J. C. Berendsen, Proteins: Struct., Funct., Genet. **17**, 412 (1993).
[30] There are five atoms for which the PDB file does not provide B-factors, presumably because it was impossible to fit the experimental data with an isotropic Gaussian of any size. All of these atoms are at the ends of lysine side chains, and have very large, highly anisotropic fluctuations in the multiple copy refinement. These atoms have been omitted when calculating the rms deviations shown in Table I.
[31] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, Jr., and P. Weiner, J. Am. Chem. Soc. **106**, 765 (1984).
[32] W. L. Jorgensen and J. Tirado-Rives, J. Am. Chem. Soc. **110**, 1657 (1988).
[33] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Clarendon, Oxford, 1987).
[34] T. Fox and P. A. Kollman, Proteins: Struct., Funct., Genet. **25**, 315 (1996).

[35] J. A. Rupley and G. Careri, Adv. Protein Chem. **41**, 37 (1991).
[36] P. J. Steinbach and B. R. Brooks, Proc. Natl. Acad. Sci. USA **90**, 9135 (1993).
[37] J. P. Ma and M. Karplus, J. Mol. Biol. **274**, 114 (1997).
[38] W. Wriggers and K. Schulten, Biophys. J. **73**, 624 (1997).
[39] D. Vitkup, G. A. Petsko, and M. Karplus, Nat. Struct. Biol. **4**, 202 (1997).
[40] H. Risken, *The Fokker-Planck Equation: Methods of Solution and Applications*, 2nd ed. (Springer, Berlin, 1989).
[41] G. Parisi, *Statistical Field Theory* (Addison–Wesley, Redwood City, CA, 1988).
[42] J. C. Smith, Q. Rev. Biophys. **24**, 227 (1991).
[43] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, J. Comput. Chem. **18**, 1463 (1997).
[44] Sparse 1.3a, copyright 1988 by Kenneth S. Kundert and the University of California. It is available from Netlib at www.netlib.org
[45] N. Gronbech-Jensen and S. Doniach, J. Comput. Chem. **15**, 997 (1994).
[46] R. Pomes and J. A. McCammon, Chem. Phys. Lett. **166**, 425 (1990).