PROTEINS: Structure, Function, and Genetics 23:97–110 (1995)

Protein Conformational Landscapes: Energy Minimization and Clustering of a Long Molecular Dynamics Trajectory

John M. Troyer[1] and Fred E. Cohen[1,2,3]

Departments of [1]Pharmaceutical Chemistry, [2]Medicine, and [3]Cellular and Molecular Pharmacology, University of California, San Francisco, California 94143-0450

ABSTRACT **Using energy minimization and cluster analysis, we have analyzed a 1020 ps molecular dynamics trajectory of solvated bovine pancreatic trypsin inhibitor. Elucidation of conformational substates in this way both illustrates the degree of conformational convergence in the simulation and reduces the structural data to a tractable subset. The relative movement of structures upon energy minimization was used to estimate the sizes of features on the protein potential energy surface. The structures were analyzed using their pairwise root-mean-square C_α deviations, which gave a global measure of conformational changes that would not be apparent by monitoring single degrees of freedom. At time scales of 0.1 ps, energy minimization detected sharp transitions between energy minima separated by 0.1 Å rms deviation. Larger conformational clusters containing these smaller minima and separated by 0.25 Å were seen at 1 ps time scales. Both of these small features of the conformational landscape were characterized by movements in loop regions associated with small, correlated backbone dihedral angle shifts. On a nanosecond time scale, the main features of the protein energy landscape were clusters separated by over 0.7 Å rms deviation, with only seven of these substates visited over the 1 ns trajectory. These substates, discernible both before and after energy minimization, differ mainly in a monotonic pivot of the loop residues 11–18 over the course of the simulation. This loop contains lysine 17, which specifically binds to trypsin in the active site. The trajectory did not return to previously visited clusters, indicating that this trajectory has not been shown to have completely sampled the conformational substates available to it. Because the apparent convergence to a single region of conformation space depends on both the time scale of observation and the size of the conformational features examined, convergence must be operationally defined within the context of the simulation.**

Key words: bovine pancreatic trypsin inhibitor, cluster analysis, conformational searching, molecular dynamics, protein tertiary structure

INTRODUCTION

Proteins show evidence of moving between multiple conformations at room temperature.[1,2] These conformational substates can exist on many levels, from large domain or hinge motions to small rearrangements of side chains.[3] Using computational techniques such as molecular dynamics (MD), we can begin to characterize the nature of these substates.[4] If we imagine the landscape of a protein's potential energy over all possible structures of that protein, then a conformational substate represents a low-energy region in this space, separated from other substates by a higher-energy barrier. If we look at a simulation in this energy landscape over time, we will observe the protein moving within the region of a substate for a time, undergoing a transition, and then moving within the region of another substate. Figure 1 schematically illustrates two conformational substates within a larger state. Multiple levels of such substates could be present on the potential energy surface.

A priori, one can assume that all structures that comprise a given substate are closer to each other in this conformational space than they are to the structures in another substate, since a potential energy barrier must be crossed to move the system to the new substate. We can therefore use the simulation to define these substates, although we can only estimate the size of these substates on the potential energy surface, because a finite-length simulation will not explore all available conformations.

Using conformational substates to analyze a MD trajectory allows for a physically reasonable interpretation of the underlying potential energy function. It is also a useful technique for data reduction

Received November 10, 1994; revision accepted March 27, 1995.

Address reprint requests to Fred E. Cohen, Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143-0450.

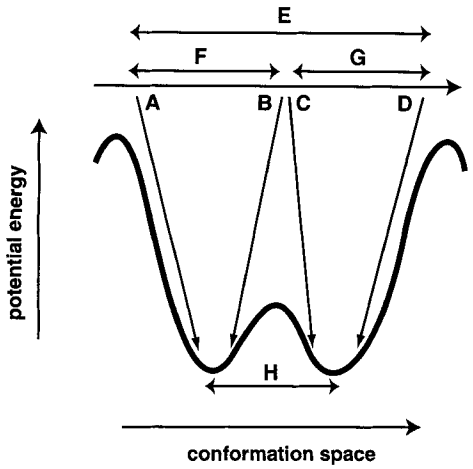


Fig. 1. A simplified model of a MD trajectory at room temperature. The *x*-axis represents an arbitrary coordinate in conformation space. The *y*-axis represents the protein potential energy. The curve represents protein potential energy minima and can be thought of as a valley floor. The MD trajectory is carried out at room temperature. Structures from the trajectory are still on the potential energy surface, but are "up the hill" from the low-energy valley floor. Structures are taken from the trajectory and energy minimized. Structures similar to each other converge within an energy well toward the same conformation (A/B, C/D). At some point in time the structures will correspond to different energy minima and will move farther away from each other (A/C). The width of energy wells can be estimated by observing both the maximum distance at which structures still converge (E, F, G) and the distance between minima (H).

and analysis of the ever larger molecular simulations, since identification of conformational substates and the transitions between them extracts meaningful information from these large data sets. In fact, the calculation of averaged properties from a simulation that consists of multiple substates is often inappropriate, since the structures are not drawn from a single, normally distributed population. For example, spatially averaged structures from MD trajectories often have large geometric distortions of mobile side chains. Even in X-ray crystal diffraction experiments, multiple structures are often better than a single, static structure at representing the calculated electron density.[5]

In this paper, we use such a strategy to analyze a 1 ns MD trajectory of the small solvated protein bovine pancreatic trypsin inhibitor (bpti).[6,7] We minimize the energy of structures generated during the simulation, moving the conformations closer to nearby local potential energy minima. The conformations, both MD-generated and energy-minimized, can then be structurally characterized through cluster analysis.

Using Energy Minimization to Elucidate Substates

Energy minimization of the MD-generated structures can be used to identify features of the energy landscape.[4] The relative movement of structures when their energy is minimized provides estimates of the size of features on the protein energy landscape. Figure 1 shows schematically how these estimates are made. Structures within energy minima will tend to move closer to each other when energy minimized (points A/B and C/D in Fig. 1). A basin of attraction can be defined as the region of conformation space in which structures converge when moved down the energy gradient. Conversely, when two structures are on opposite sides of a potential energy barrier, they will diverge upon energy minimization (points B/C in Fig. 1). Finally, when the structures are farther apart, their relative movement becomes uncorrelated with their position on the potential energy landscape (points A/D in Fig. 1).

Multiple levels of substates can be accessible to proteins at room temperature.[1,2] The presence of a substate of a certain size is operationally defined when neighboring structures in an MD trajectory stop converging upon minimization and begin to diverge. Two separate estimates of the width of these features can be made. The first estimate gives the sizes of basins of attraction, defined by the maximum distance at which structures still converge. Distances E and F/G illustrate this estimate for two levels of substates on the energy surface in Figure 1. The second estimate (distance H in Fig. 1) is of the distance between clusters after minimization. These two estimates should give a consistent picture of the magnitude of features on the protein energy landscape. This technique provides more information on the minima of the potential energy surface than on the energy barriers between these minima. It is important to recognize that the free energies associated with these substates and the transitions between them are accessible only if enough barrier crossings are observed in the simulation.

Clustering

Since we assume that similar protein structures define a conformational substate, we will cluster both the structures generated in the simulation and those structures after energy minimization. Clustering conformations of a large system such as a protein is useful, since it offers a way to characterize a large amount of data and guides the analysis to focus on the most important changes seen in the course of the simulation.

Although clustering techniques have been applied to the molecular dynamics trajectories of several short peptides,[8,9] we are not aware of similar analyses for protein simulations. Small molecules have often been structurally classified using clustering algorithms.[10–13] These capabilities are becoming available in general MD analysis tools such as Scarecrow,[14] MacroModel/XCluster,[15] and MD Toolchest (Ravishankar and Beveridge, personal communication). Shenkin and McDonald have developed a graphic tool to carry out analyses similar

to those in this paper. They report the results of clustering the structures generated by MD for several small molecules, although they did not carry out any energy minimization.[15]

There are presumably several levels of substates available to proteins.[1,2] Using a hierarchical clustering algorithm, it should be possible to distinguish both the substates and the hierarchy of states within other states defined by increasing distance criteria. To enumerate both the very small substates and the very large substates that contain them, we would need to sample the entire time frame at the highest level of detail, which is computationally infeasible. Instead, we take representative intervals at orders of magnitude from 1 ps to 1 ns and use them to characterize the general features of the bpti conformational landscape.

To cluster protein structures, a single measure of their relatedness is needed. We use root-mean-square (rms) Cartesian deviations of the C_α atoms of superimposed proteins to estimate their similarity:

$$\text{rms Cartesian deviation} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|x_i - Tx_i'|^2}$$

where n is the number of atoms, x_i is the Cartesian coordinate of the ith atom, and T is the transformation matrix that best superimposes the two structures. The rms Cartesian deviation condenses the conformational dissimilarity between two molecules into a single representative number. It is a commonly-used, well-understood global measure of structural similarity for proteins close in structure.[16]

The rms deviation between two structures is a measure of their differences. We can imagine a "conformational space" in which the rms deviation between two structures represents the Euclidean distance between them. This representation is slightly different than the energy landscape illustrated in Figure 1, since potential energy is not involved in this projection. We can use the rms deviation matrix $\{rms_{ij}\}$, where i and j are any two structures, to create a set of points which represent the disposition of structures in this multidimensional Cartesian space.[9,17–19] We can then visualize the trajectory as a path across the space, using projections onto the most important two or three dimensions. Energy minimization causes relative changes in position of the points in this landscape, and clusters of structures will also cluster in this representation.

MATERIALS AND METHODS

Molecular Dynamics

We have analyzed a molecular dynamics simulation of bovine pancreatic trypsin inhibitor (bpti) carried out by Guenot.[6,7] The details of this simulation will be presented in more detail elsewhere. Briefly, the simulation began with the structure 4pti[20] from the Brookhaven Protein Data Bank. Hydrogen atoms were added to the protein structure with the AMBER edit module. 280 TIP3P water molecules, taken from a box of water molecules equilibrated at 298 K, were added in a 5 Å shell around the protein. The rationale for this solvation model and evidence of its utility have been discussed previously.[7,21] The total bpti system consisted of 1746 atoms. The system was carefully relaxed before the start of the simulation in order to prevent instabilities in the solvent shell. The water molecules were energy minimized and 10 ps of MD were carried out with the protein coordinates fixed. The entire system was then energy minimized to 0.1 kcal/mol Å. To allow the water molecules to relax without distorting the protein structure, separate minimizations were carried out, first with harmonic position restraints on the protein backbone of 100, 50, 15, and 2 kcal/mol Å^2, and finally with no position restraints.[7,21] In general, one or two water molecules escaped from the "droplet" during the 1 ns simulation. The MD simulations were carried out with the program AMBER 4,[22] starting with the minimized structure described above. The simulations used all-atom models and a distance dependent dielectric function, with $\epsilon = 1.0r$, and an 8 Å nonbonded cutoff distance. The simulations were carried out coupled to a temperature bath at 298 K with a temperature coupling constant of 0.2 ps^{-1}. The MD step size was 0.0015 ps, and the SHAKE algorithm was used to constrain all bond lengths. The nonbonded list was updated every 10 steps. The simulation was carried out to 1020 ps. Details of the simulation[6] and its extension to over 15 ns will be published elsewhere (Guenot and Kollman, in preparation).

The electrostatic model and solvent shell used in this calculation can reproduce the behavior of protein simulations with more complete solvation models.[7,21,23] It is likely that the limitations of this model have introduced artifacts into this trajectory, especially in the behavior of charged side chains on the protein surface. However, we are more interested in developing an analysis method and in the features of the potential energy surface at the structural level of the C_α atoms. We expect that the general conclusions of our analysis would hold for simulations with a more complete solvent model.

Energy Minimization

An energy minimization protocol was applied to sets of structures spanning intervals of 1, 10, 102, and 1020 ps. Energy minimization was carried out with the AMBER minmd module. For all calculations, one hundred cycles of steepest descent minimization were carried out, followed by conjugate gradient minimization. The minimizations were stopped when the norm of the potential gradient changed less than a given amount in a single step (standard AMBER minimization). The complete system, includ-

ing the shell of water molecules, was used in the calculation.

When a gradient convergence criterion of 0.1 kcal/mol Å was used to define convergence to a local minimum, the resulting structures failed to show any clustering at the 1 and 10 ps time scales. Clusters began to appear in both the 1 and 10 ps intervals when the structures were energy minimized again to a gradient convergence criterion of 0.01 kcal/mol Å. In order to test if further minimization would be fruitful, the 1 ps- interval was minimized again to a gradient convergence criterion of 0.001 kcal/mol Å. The maximum gradient of the converged structures ranged from 0.008–0.025 kcal/mol Å, although the structures at 0.12 and 0.45–0.49 ps failed to converge within 99,000 steps of conjugate gradient minimization. Other structures required 10,000–35,000 steps to meet the convergence criteria. The 10, 102, and 1020 ps trajectories were all minimized with a 0.01 kcal/mol Å convergence criterion, and required up to 5,000 steps of minimization to converge.

Since approximately 10,000 steps of minimization required 1 CPU hour on a Hewlett-Packard 9000/735 workstation, each structure took 0.5–3.5 h of computer time to energy minimize. The minimization of every saved structure was not computational feasible. Therefore, several representative structures were drawn at uniform intervals from from each of the segments of the trajectory and energy minimized: 100 were taken from the 1 ps segment, 50 from the 10 ps segment, and 85 from each of the 102 and 1020 ps segments.

The smaller intervals were generated from the larger trajectory as follows. Several 1 ps intervals were analyzed, although the one described in detail here was taken from the 102–103 ps interval of the original 1020 ps trajectory. When we began this study, we had only the coordinates of the trajectory stored at 0.15 ps intervals without velocities. Since it proved necessary to examine structures at a finer level of detail, we generated a 10 ps interval as the final portion of a new 20 ps trajectory starting from the 102 ps structure with a random distribution of velocities corresponding to a Boltzmann distribution at 298 K. The 102 ps trajectory was taken from the initial portion of the 1020 ps trajectory.

Clustering

In an effort to explore the extent of conformational sampling across the different time scales, the protein structures were clustered. Each protein structure was superimposed on every other structure in that trajectory using the method of Kabsch.[24] Superpositions were based on the optimal least-squares overlap of equivalent C_α atoms. The rms deviation of the Cartesian coordinates of C_α atoms and of all heavy atoms were measured. Only the results for C_α rms deviation are presented; all atom rms deviations gave similar results.

The pairwise rms deviations were used to build a matrix. Each element i,j of this symmetric matrix $\{rms_{ij}\}$ is the comparison of the two structures i and j separated in time along the same trajectory. The rms deviation matrices retain the original time-ordering of the structures from the trajectories.

The clustering defined by the rms deviations was first evaluated subjectively. The matrices were contoured, color-coded, and displayed for a visual evaluation of the conformational clustering of the trajectories. A cluster here is loosely defined as a group of structures that has mutually low rms deviations with respect to each other.

A second, more objective analysis was carried out using a hierarchical clustering method. Hierarchical clustering methods first associate close structures, and then gradually combine smaller clusters as a distance criterion is raised. At each step of the algorithm, the distance between clusters was taken as the average of the distances between the points in one cluster and the points in the other. Other criteria can be used, such as "single linkage," which uses the shortest distance between clusters as the combining criterion, or "double linkage," which uses the longest distance. However, the clusters defined by these criteria did not agree with those from a visual analysis as well as did the clusters defined by the averaged distance criterion. The clustering and other statistical analyses were performed using S-PLUS.[25]

Multidimensional Scaling

Multidimensional scaling is an analysis technique developed in the behavioral and social sciences that is closely related to principal component analysis.[26] In multidimensional scaling, only the pairwise dissimilarities δ_{ij} among a set of I points are known. The procedure treats these dissimilarities as Euclidean distances in a K-dimensional space, and solves for the set of Cartesian coordinates x_{ik} in this space that would best reproduce the distances. This is mathematically equivalent to the principal component projection of a set of data points, except that we do not start with the $I \times K$ matrix of Cartesian coordinates, but the $I \times I$ matrix of distances between them.

If Δ is the $I \times J$ matrix of pairwise distances δ_{ij}, then a double centered matrix Δ^* can be constructed of elements δ_{ij}^* such that

$$\delta_{ij}^* = -\frac{1}{2}(\delta_{ij}^2 - \delta_{i\cdot}^2 - \delta_{\cdot j}^2 + \delta_{\cdot\cdot}^2)$$

$$\delta_{i\cdot}^2 = \frac{1}{J}\sum_j \delta_{ij}^2$$

$$\delta_{\cdot j}^2 = \frac{1}{I}\sum_j \delta_{ij}^2$$

$$\delta_{..}^2 = \frac{1}{IJ}\sum_i\sum_j \delta_{ij}^2.$$

In our case, Δ and Δ^* are both square, symmetric matrices, and therefore $I = J$. It can be shown that each element δ_{ij}^* of Δ^* is equivalent to $\Sigma_k x_{ik}x_{jk}$, or in matrix notation $\Delta^* = XX^T$, where X is the $I \times K$ matrix of Cartesian coordinates. The analysis now proceeds exactly as in principal component analysis, since Δ^* corresponds to the correlation matrix used in principal component analysis. Using matrix algebra, we can solve $\Delta^* = B\Lambda B^T$, where B is the matrix of eigenvectors and Λ is the diagonal matrix of eigenvalues. The coordinates in a particular dimension k are therefore the kth column of the eigenvector matrix scaled so that their sum of squares is equal to the corresponding eigenvalue.

RESULTS

Relative Movement During Energy Minimization

Figure 2 shows the rms deviations of C_α atoms before and after energy minimization over all time scales studied. In general, rms deviation between pairs of structures increased with the time interval between them. The changes in potential energy between the bpti MD structures and their energy-minimized counterparts were on the order of 500 kcal/mol; their C_α rms deviations were between 0.2 and 0.4 Å. Only small geometric rearrangements were necessary to reach these nearby local minima on the potential energy surface. The volume of the protein decreased by approximately 2% during energy minimization. No obvious energy trends were observed along the time course of the trajectories or its energy minimized equivalent.

Points above the slope 1 line of Figure 2 indicate pairs of structures that have moved farther apart on energy minimization. Most pairs of structures in Figure 2 have moved closer together, but there are two groups of points that move farther apart: pairs of structures with very short time separations which diverge with approximately 0.25 Å rms deviation, and pairs of structures greater than 200 ps apart apart in the trajectory which diverge with greater than 0.5–0.8 Å rms deviation between them. These two features correspond to the two levels of energy minima. As the time interval increases, structures are no longer in adjacent minima, and their chances of converging or diverging become less correlated with their separation in conformation space.

Minimization and Clustering of the 1 ps Trajectory

Some pairs of structures at very short time scales move apart when energy minimized, indicating they are separated by a potential energy barrier. Figure 3A illustrates the results of energy minimization of the 1 ps MD trajectory as a color-coded matrix of C_α atom rms deviations. The left half of this matrix shows the rms deviation between structures in the unminimized trajectory. In this trajectory, structures close in time are also close in rms deviation, and structures move farther apart smoothly with time. All 1 ps intervals observed appear similar to this figure.

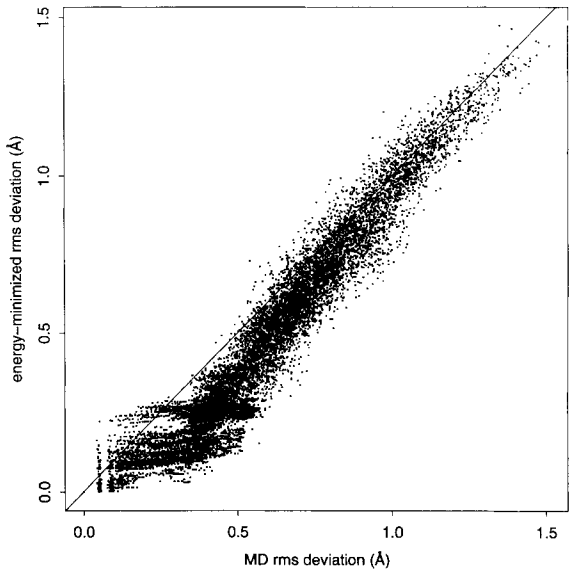
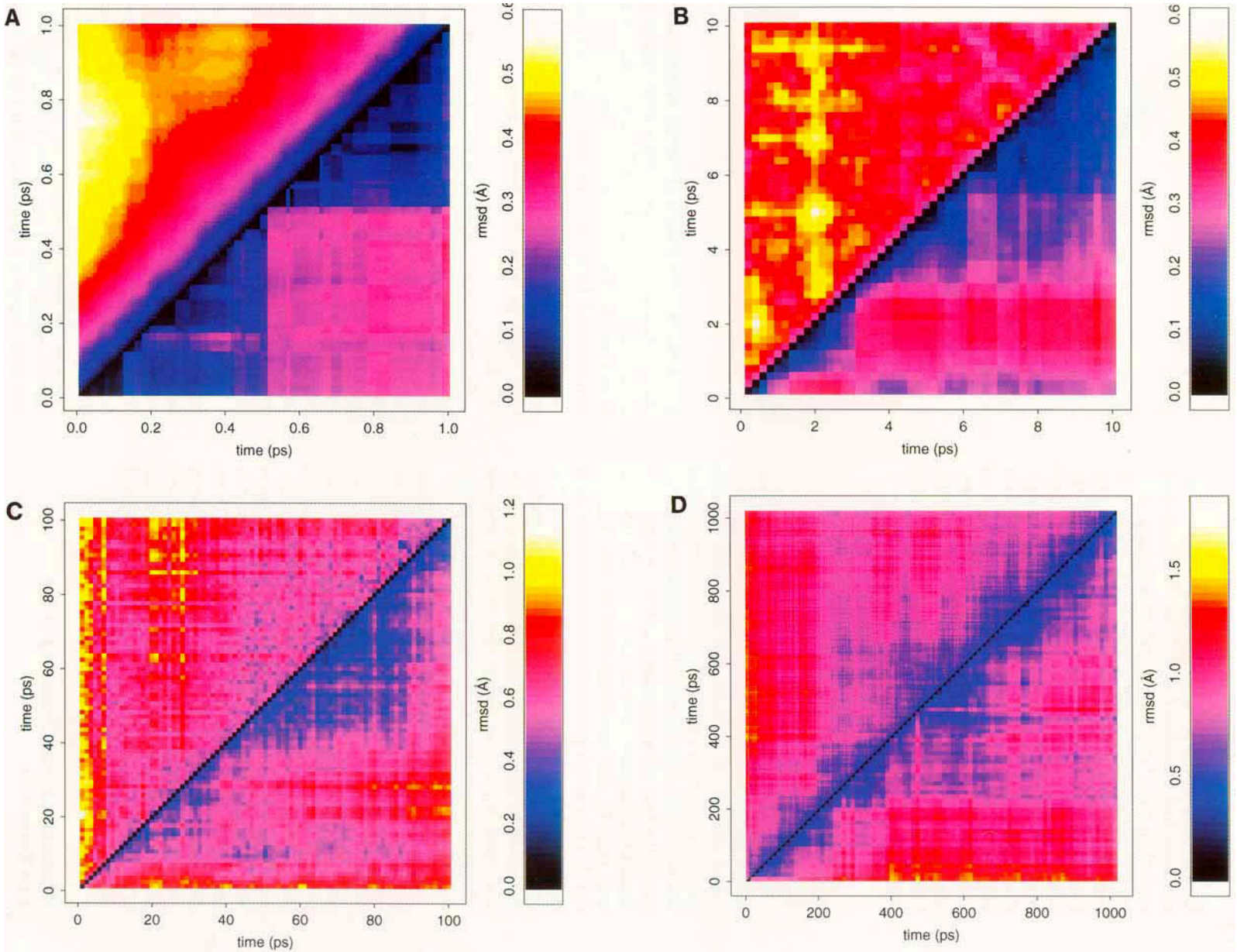


Fig. 2. C_α rms deviation values for each pairwise comparison of MD structures plotted against the C_α rms deviation values for the equivalent minimized structures. A line with a slope of 1 is also displayed.

The right half of Figure 3A shows the rms deviations between the same structures after energy minimization. The relationships between these structures are very different. The largest rms deviation changes from 0.57 Å in the MD trajectory to 0.36 Å in the minimized sequence. Most of the structures move closer together upon minimization, indicating that the structures are within a single energy well. Structures that are similar and contiguous in time are displayed as low-rms deviation squares along the diagonal.

Two levels of substates can be seen: two large clusters dominate the map, with many smaller clusters inside of them. The two larger clusters have a well-defined transition between them after 0.51 ps. (The transition falls in the middle of this interval only by chance.) The structures differ within each of the clusters by less than 0.17 Å rms deviation. Deviations between structures in the two different clusters range from 0.22 to 0.35 Å. Smaller subclusters with transition times around 0.1 ps can be seen as dark blue/black squares along the diagonal within the two larger clusters. Intracluster differences for these smaller subclusters range from 0.001


A
time (ps)
time (ps)
rmsd (Å)
B
time (ps)
time (ps)
rmsd (Å)
C
time (ps)
time (ps)
rmsd (Å)
D
time (ps)
time (ps)
rmsd (Å)

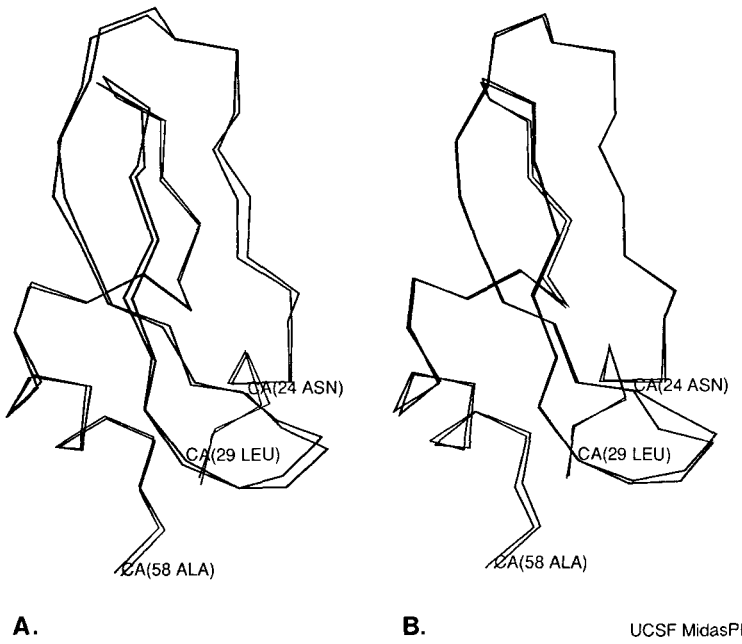


Fig. 4. **A:** Unminimized C_α chains from the 0.19 and 0.83 ps structures of the 1ps bpti MD trajectory. The chains have an rms deviation of 0.46 Å. **B:** The same two structures after energy minimization. The chains now have an rms deviation of 0.31 Å. Differences are predominantly in the C-terminus and the loop consisting of residues 24–29.

to 0.05 Å rms deviation, with intercluster differences about 0.1 Å rms deviation.

Of the 4950 rms deviation comparisons of the matrix shown in Figure 3A, 396 increase on energy minimization, indicating that they lie on opposite sides of an energy barrier. The pairs of structures that move farther apart are close in time on the boundaries between conformational clusters. The structures at 0.41–0.51 ps, which immediately precede the main transition, move away from the structures at 0.52–0.59 ps, immediately after this transition. The pairs of structures on the boundaries of the smaller subclusters also move apart, indicating that the larger clusters do not consist of a simple energy minimum, but instead are composed of multiple smaller minima. The pairs of structures in the 0.1 ps subclusters do converge, sometimes to less than 0.001 Å rms deviation, indicating that they lie in a single energy minimum.

Fig. 3. Color-contoured C_α rms deviation matrices. Upper left triangles are from the original trajectories. Lower right triangles are from the corresponding energy-minimized structures. **A:** 100 structures from the bpti 1 ps MD trajectory. The blue and black squares, contiguous regions of mutually low rms deviation, signify conformational clustering. These clusters are presumably in basins of attraction associated with different energy minima. **B:** 50 structures taken from the 10 ps MD bpti trajectory. **C:** 85 structures taken from the 102 ps MD bpti trajectory. **D:** 850 structures taken from the 1020 ps MD bpti trajectory, and 85 energy-minimized structures.

Structural Significance of Short Time Scale Clusters

Representative structures from both of the large clusters of the 1 ps trajectory are shown in Figure 4. Structures at 0.28 and 0.84 ps are displayed both from before (Fig. 4A) and after (Fig. 4B) energy minimization. The unminimized structures have a rms deviation of 0.46 Å, and this difference is distributed throughout the chain. The energy-minimized structures have an rms deviation of 0.31 Å, but this difference is localized in several regions. The two clusters differ predominantly in small ($< 15°$) correlated dihedral angle changes in the C-terminus and in the loop formed by residues 24–29. No differences in hydrogen bonding are seen between the two clusters. The smaller subclusters in this 1 ps interval differ by smaller dihedral angle changes in the same regions.

Minimization and Clustering of the Longer Trajectories

The MD and energy minimized structures of the 10 and 100 ps intervals are shown in Figure 3B, C. In this particular 10 ps interval (Fig. 3B), the first 3 ps and the last 7 ps form clusters. Unlike in the other cases we examined, here the clusters are correlated with a change in energy: the minimized potential energy of the protein in the first 3 ps is approximately −720 kcal/mol, which falls to −729 kcal/mol at the transition and slowly rises to −720

kcal/mol again by the end of this 10 ps interval. Although the transition is clear in the original MD structures, the clustering is apparent only after energy minimization.

Figure 3C shows the rms deviations for both sets of 100 ps intervals. Energy minimization also helps elucidate the clusters seen in this interval. The protein remains in a single cluster from 40 to 90 ps.

The C_α rms deviations for the complete 1020 ps trajectory and the corresponding energy-minimized structures are shown in Figure 3D. Although only 85 structures were energy-minimized from the 1020 ps trajectory (one every 12 ps), we have shown 850 unminimized structures in the upper half of Figure 3D to demonstrate that the protein is not undergoing large fluctuations in the 12 ps intervals between the 85 frames. Of the 3570 comparisons between the energy minimized structures, 856 pairs of structures move farther apart upon energy minimization.

Using a hierarchical clustering algorithm, seven clusters can be distinguished in both the 1 ns MD trajectory and its energy-minimized counterpart at a 0.65–0.7 Å level. The clusters detected in the MD structures and the energy-minimized structures are very similar; they also agree with those determined from visual inspection of the matrix. The hierarchical clustering of the minimized structures is shown in Figure 5. The clusters consist of structures contiguous in time, and the boundaries vary by only 1–2 structures (12–24 ps) between the MD and energy-minimized clusters. These energy-minimized clusters are over the intervals 24–96, 108–240, 252–384, 408–648, 660–816, 828–948, and 960–1020 ps. Since the structures were taken from the trajectory every 12 ps, the transitions cannot be further localized. Intracluster differences are usually less than 0.5 Å rms deviation, while intercluster differences range from 0.7 to 1.5 Å. In addition, the trajectory does not return to previously-visited clusters. Finally, three structures are not included in the other clusters at the 0.65 Å rms deviation level. The structure at 12 ps is an outlier, close to the starting X-ray structure, and is generated before the trajectory moves away from its crystal conformation. The structure at 396 ps is at the transition between clusters and is an outlier in both the MD and energy-minimized sets. The structure at 480 ps, on the other hand, has a much higher energy than the conformations preceding and following it. It has become "stuck" in a high energy minimum during the energy minimization process, and is far from its neighboring conformers as measured by rms deviation. This structure at 480 ps does not appear anomalous in the unminimized MD trajectory.

Structural Significance of the Long Time Scale Clusters

The averaged structures from each of the seven clusters are shown in Figure 6. Although there are differences distributed throughout the chain, the clusters differ primarily in the conformation of the loop consisting of the residues 11–18. The loop pivots around the disulfide bridge at residues 14 and 38 through an arc of approximately 30° during the 1020 ps of the trajectory. Residues 13 and 17, the "corners" of this loop, show the greatest displacements. The C_α atoms of these residues move by 3.2 and 4.2 Å, respectively. The chain termini also change their orientation throughout the trajectory.

The hydrogen bonding patterns of these averaged structures were examined to see if the loss and gain of hydrogen bonds were directly driving the conformational change of loop 11–18. The "corners" of this loop extend into the solvation shell and do not make intraprotein hydrogen bonds. Since the definition of the clusters is primarily associated with motion of this loop, transitions between clusters are not associated in a simple fashion with the change of 1 or 2 hydrogen bonds. However, the hydrogen bond Ile-18 N–Tyr-35 O, located at one end of this loop, is lost over the course of the calculation. The hydrogen bond Cys-38 N–Cys-14 O between the cystines at the pivot point is created over the simulation, as is the side chain hydrogen bond 32-Thr OG–21-Tyr OH. This loop also has highly variable conformations in different crystal structures of bpti, although its range of motion seen in this MD trajectory is much larger than the range seen in various crystal forms.[27]

DISCUSSION

Dependence on the Force Field

The MD trajectories analyzed here will be discussed in detail elsewhere (Guenot and Kollman, in preparation). We use this particular simulation only as an example and wish to highlight several useful techniques for the analysis of large conformational data sets. This solvation model does produce trajectories that are in good agreement with the crystal structure, even for protein structures that are unstable in vacuum simulations.[7,21] We should note, however, the limitations of the force field that was used in this simulation, particularly the electrostatic and water model.

Arnold and Ornstein used a similar solvation model in a molecular dynamics simulation of bacteriophage T4 lysozyme.[23] They found that a 6 Å shell of water around the protein behaved nearly as well as a 10 Å shell. The criteria for judging the simulations were based on agreement with the crystal structure coordinates temperature factors, small temperature difference between protein and solvent, and atomic fluctuations sufficient to sample conformation space. Some regions of the protein in the Arnold and Ornstein simulation became unsolvated when water molecules in the 6 Å solvent shell became clustered. This did not occur with the solvent shell of the Guenot and Kollman simulation.

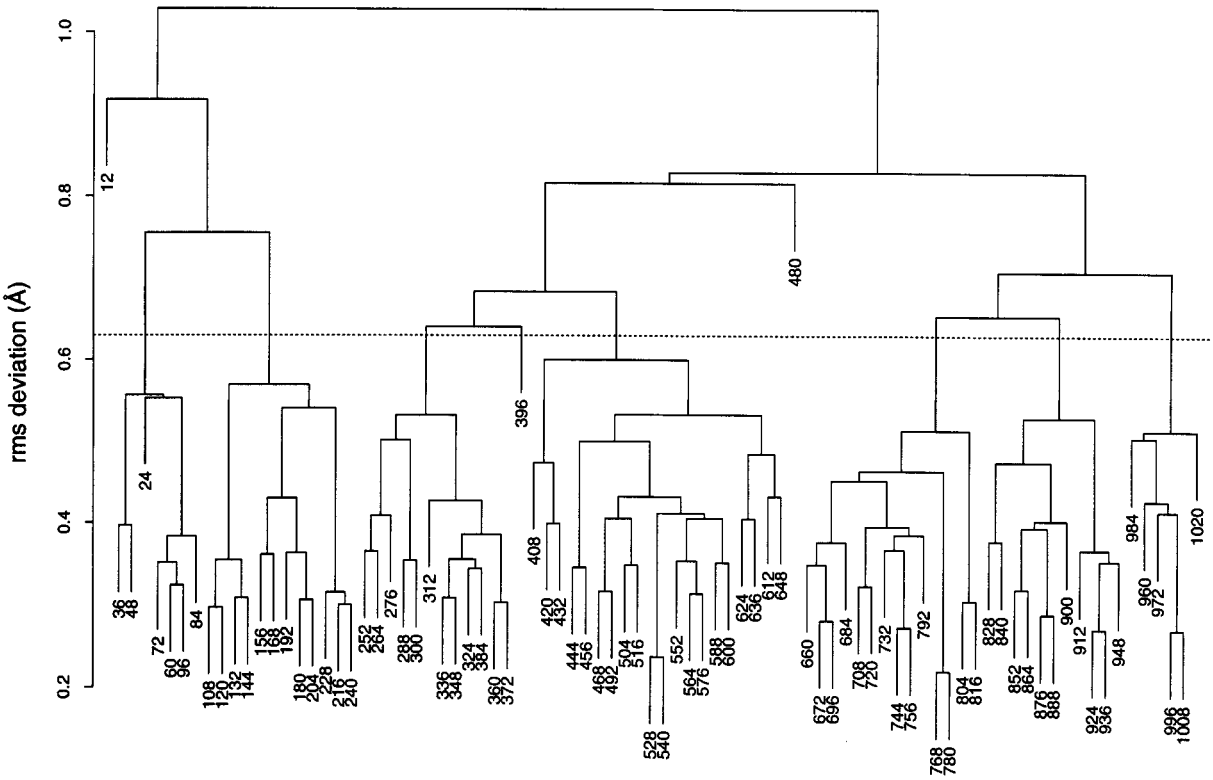


Fig. 5. Hierarchical cluster diagram of the 85 energy-minimized structures from the 1020 ps MD bpti trajectory using C_α rms deviation. A 6.5 Å line separates the structures into 7 clusters.

Fig. 6. Averaged structures of the seven clusters present in the 1020 ps trajectory. The view is approximately from the top of the orientation shown in Fig. 3A. The structures are positioned to emphasize the loop region around residues 11–18, which pivots by approximately 30° over the course of the trajectory. The chains in this region are colored in their order from the trajectory: red, orange, yellow, green, blue, indigo, violet.

The distance-dependent dielectric function in the Guenot simulations will cause the damping of long-range electrostatic forces. Combined with the decrease in radius of gyration and the surface tension and other edge effects caused by the 4 Å water shell, this may inhibit large motions in the protein. This would, in turn, affect the transition times of the protein moving between conformational substates, such as the swing of the loop seen in this simulation. The sizes of these substates are measured by rms deviation in this study. These sizes are more likely to depend on steric volume exclusion and the bonded geometry of the protein than the solvent model used.

Conformational Substates in bpti

From the relative movement of structures upon energy minimization, as is shown in Figure 2, we see that the backbone energy landscape of bpti was dominated by features of two sizes. At short time scales less than 1 ps, small features of the conformational landscape were seen separated by 0.1 and 0.25 Å rms deviation. These resulted from small displacements within localized loop regions that can be attributed to correlated changes in backbone dihedral angles. Since all structures in this interval move closer together upon energy minimization, we estimate the size of these features only from the distance between clusters (distance H in Fig. 1).

As is illustrated in Figure 1, we have two estimates of the size of the larger conformational substates. As seen in Figure 2, structures stop converging upon minimization at a separation of 0.5–0.8 Å rms deviation. This defines a basin of convergence. After minimization, adjacent clusters are separated by 0.7–1.0 Å rms deviation (see Fig. 3D). Both of these estimates can be used to produce a consistent picture of the distance between energy minima. Seven of the larger conformational substates were observed during the 1 ns MD trajectory, and they differed primarily in the conformation of the residue 11–18 loop region.

The underlying structure of the potential energy landscape at the 1 ps time scale was not apparent from an examination of the MD trajectory itself. The kinetic energy of the dynamic trajectory masked these very small features, while after energy minimization they were clearly delineated. However, the clustering of the full 1 ns trajectory did not change significantly with energy minimization. At this time scale, the conformational substates observed were far apart, and the movement to smaller nearby energy minima did not affect the overall disposition of the structures. Energy minimization at this level of conformational difference still gives information about the potential energy landscape, but does not change the overall picture of structural transitions.

Projection onto the Conformational Landscape

The rms deviations between structures are represented as distances between points in a Cartesian space in Figures 7 and 8. No potential energies are shown in these figures; the disposition of points relates only to the differences in rms deviation between structures; thus the trajectory can be shown as a path in this conformational space. The projections of the 1 ps interval into the planes defined by the largest 3 eigenvalues are shown shown in Figure 7 along with the the magnitudes of the 10 largest eigenvalues A combined rms deviation matrix was constructed for both the 100 MD-generated structures and the 100 energy-minimized structures. This 200 × 200 matrix was then used to project the two 100-point trajectories onto a Cartesian space. The 1 ps MD-generated trajectory can be seen as a smooth pathway through this space. When these structures are energy-minimized, they move closer together into two smaller clusters, with even smaller clusters inside. In Figure 8 the 85 structures of the 1 ns MD trajectory and their energy-minimized equivalents were projected separately. The clusters defined in Figure 5 can be discerned, although the separations between clusters are more clear in three dimensions.

Two- and three-dimensional projections of rms deviations must be carefully interpreted. This is a projection of a high-dimensional set of points onto a low-dimensional space, and distortions in the relationships between the points can result. In general, the projected distances are smaller than the original distance matrix, and the projections are very sensitive to the largest distances between points. A projection of the 240–1020 ps portion of the 1 ns trajectory looks quite different than Figure 8, since the extent of the first dimension, which was devoted to representing the large distances between the first cluster and the rest of the trajectory, is now used to better represent the distances between the structures in the final portion.

More importantly, the volume described by rms deviation cannot be adequately described with low dimensionality. Structures taken from an MD trajectory far apart in time tend to have similar rms deviations. MD trajectories are often monitored by rms deviation from a reference structure, and show a characteristic plateau with time. The $\{rms_{ij}\}$ matrices of these trajectories describe mutually equidistant points, which leads to substantial distortions when projected into a few dimensions. This imparts a characteristic spiral shape that is seen in projections of rms deviations at all time scales. The curve made by the unminimized trajectory in Figure 7 is an example. These projections are shown only as an alternative method of visualizing the clusters defined by the rms deviation matrices. It is difficult to associate specific structural characteristics with the projected degrees of freedom.

Other Estimates of Conformational Substates

Noguti and Go examined the energy landscape and the substates of bpti in vacuo.[28] They used collective variable normal mode Monte Carlo to simulate the fluctuations of the protein. The simulation consisted of 5×10^5 Monte Carlo steps, which the authors estimate corresponds to a time interval 2.5–25 ns. Two types of motions were observed and differentiated by their magnitude and their extent of localization in the protein structure. One class of motions involved 0.2–0.4 Å displacements, and corresponded to harmonic motions of local regions within an energy well. Fluctuations larger than this

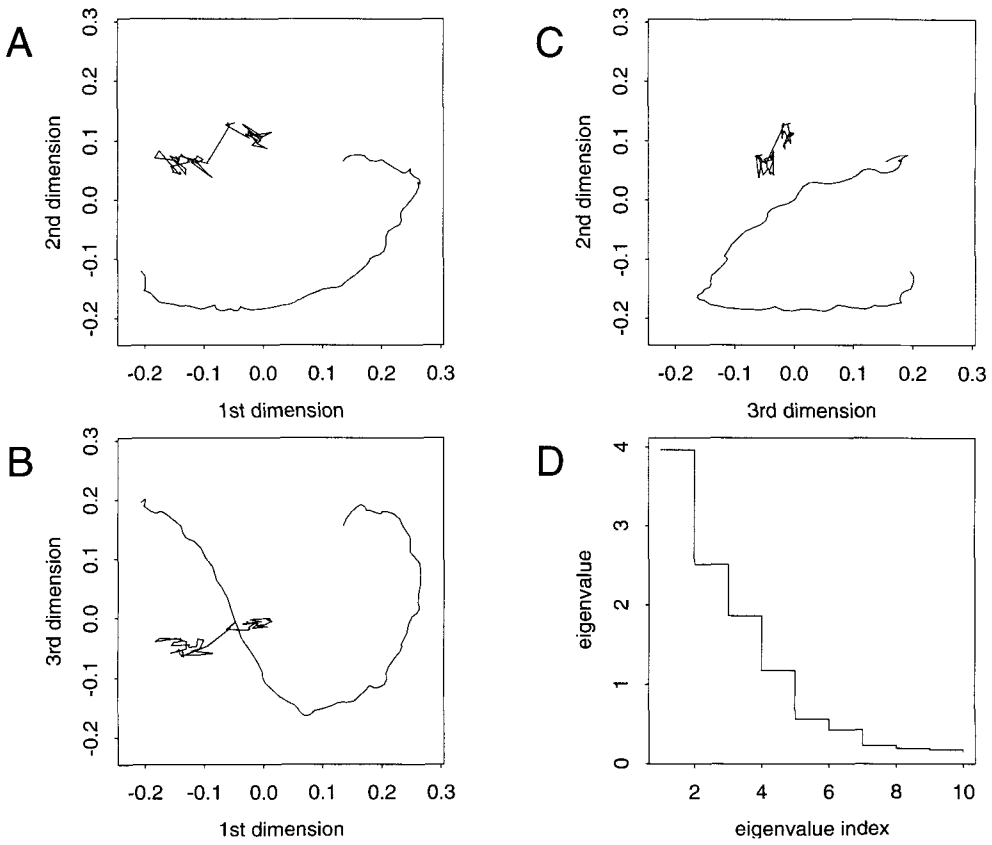


Fig. 7. **A, B, C:** Projections of the unminimized and energy minimized 1 ps MD trajectory on the planes defined by the 3 largest eigenvalues of the projection of their combined C_α rms deviation matrix. **D:** The magnitudes of the 10 largest eigenvalues for this projection.

were attributed to movement between wells, and tended to involve a more global change in shape of the molecule. These studies used a quite different force field than the ones described in this manuscript, since dihedral angles were the only degrees of freedom allowed to vary in the model. Despite the differences between the Guenot and Kollman and the Noguti and Go simulations, there is a good correspondence between estimates of the size of wells on the two potential energy surfaces.

Elber and Karplus performed an analysis similar to ours on a series of myoglobin structures generated from an MD simulation in vacuo.[4,29] They compared the rms deviation of pairs of structures before and after energy minimization. Elber and Karplus found that after 0.15 ps, the trajectory had crossed into another energy minimum with a 0.4 Å rms deviation from the initial minimum. This difference is larger in magnitude than the minima found in this paper (0.4 Å at 0.15 ps in their study vs. 0.1 Å rms at 0.1 ps and 0.25 Å rms at 0.5 ps in our study). They did not analyze the structural differences between the two myoglobin minima, and only compared seven minimized structures. In addition, only one solvent molecule was included in the Levy et al. simulation.[29] The absence of solvent in this simulation may have affected the magnitude of the protein fluctuations during the simulation. The precise details of the equilibration methodology could also impact the structural comparisons. Alternatively, the highly disulfide-bridged bpti might be expected to be more rigid than myoglobin, having more restricted energy minima and moving between these minima more slowly. Finally, myoglobin has 153 residues, while bpti has 58. Since rms deviations tend to be proportional to chain length, the deviations between myoglobin structures would be expected to be higher.[16]

Amadei et al. analyzed MD trajectories of hen egg white lysozyme in vacuo and in solution.[18] The trajectories were approximately 1 ns long. Utilizing atomic correlations along the trajectory, they also found two types of motion. An "essential" subset, consisting of less than 1% of the degrees of freedom in the molecule, accounted for most of the motion in the simulation. These essential degrees of freedom are often hinge-bending or other domain motions. In the current study, this would correspond primarily

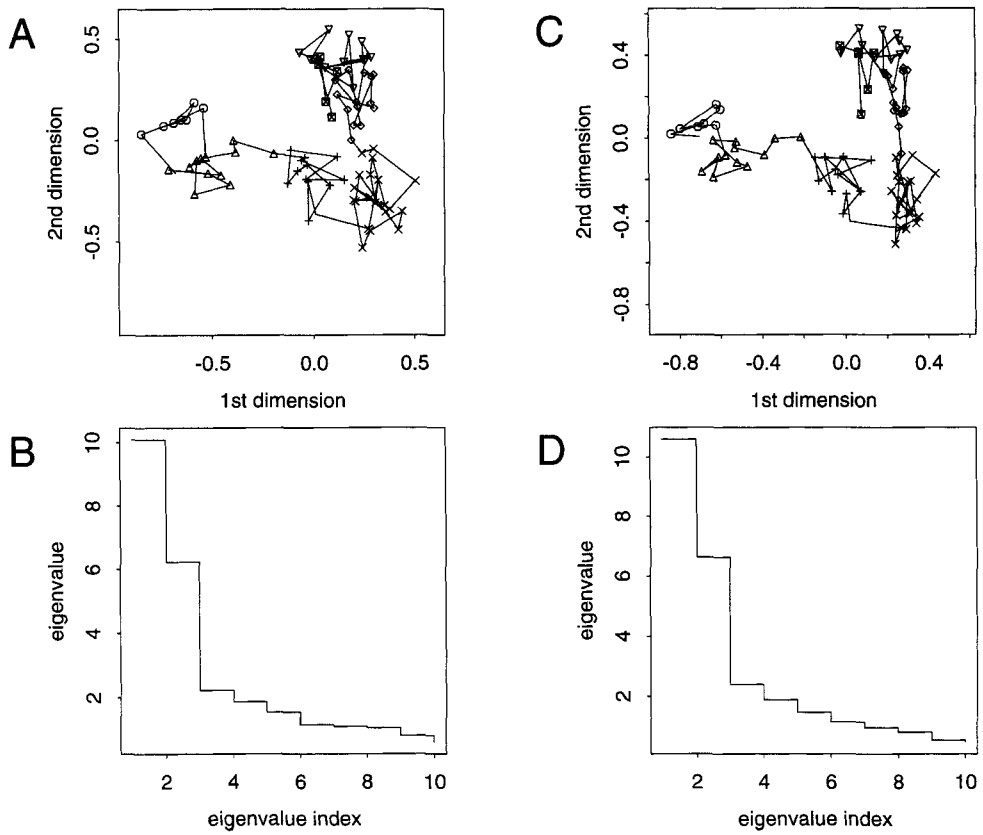


Fig. 8. **A:** Projection of 85 unminimized structures from the 1 ns MD trajectory on the planes defined by the 2 largest eigenvalues. The trajectory begins on the left side of this figure. The 7 clusters identified in Figure 7 are identified with different symbols. **B:** Projection of the same 85 structures after energy minimization, using the same symbols. **C:** The magnitudes of the 10 largest eigenvalues for the projection of the unminimized structures. **D:** The magnitudes of the 10 largest eigenvalues for the projection of the minimized structures.

to the pivoting motion of the loop of residues 11–18, and to a lesser extent rigid body motions of the other loops. The other degrees of freedom of the protein are characterized as "irrelevant" independent Gaussian fluctuations.

Although the essential dynamics technique has broader applications, in this initial study by Amadei and co-workers it is essentially used as an analysis technique. Essential dynamics emphasizes continuous movement along the largest, slowest modes of motion in a simulation. In contrast, the very simple rms deviation matrix gives a picture of protein motion dominated by static conformational substates with relatively sharp transitions between them. Both pictures of protein dynamics should prove helpful. In addition, the projection along these essential degrees of freedom has a structural interpretation, unlike the projections in this study. However, the long time scale distances in the essential dynamics projections apparently still are distorted in the projections. This can be seen by the characteristic spiral-shaped curve in Figure 9 of Amadei et al.[18]

Conformational Sampling

How do we know when an MD trajectory has explored the appropriate portions of conformation space, and sampled these regions adequately? The stabilization of potential energy over the course of a simulation, often used as evidence of "equilibration," obviously does not correspond to equilibration in the conformational sense. Although individual degrees of freedom may completely sample each available energy minima many times over the simulation, cluster analysis shows that the overall structure does not. In the 1 ns trajectory, this protein explores seven different ensembles, each with at least a 0.7 Å rms deviation from the others. Since the protein does not return to previously visited clusters, we can not assume that we have completely sampled all the available conformations of this region of the protein. Calculations that rely on complete sampling of this region may be affected. For example, calculation of the free energy difference between the wild-type protein and a protein with a

mutation in this loop may depend on complete sampling of the loop.

Trajectories which appear to have converged to a single substate at one time scale are seen to move away over longer periods of time in these simulations. Since the multiple minima of the potential energy surface cause clusters to be seen at several scales, apparent convergence to a particular region of conformation space is dependent on the time scale of observation. Figure 3C illustrates, for instance, that if only the first 100 ps of the simulation is considered, the trajectory appears to converge to a single region of conformation space. It is only in the context of the longer simulation shown in Figure 3D that we realize other conformers are accessible to the protein. Evidence of a larger conformational movement can also be seen in Figures 3D and 7. The structures from 24 to 240 ps, which make up the first two large clusters of the nanosecond trajectory, consistently move away from the rest of the trajectory when energy-minimized and have higher rms deviations with respect to the later structures. The protein appears to have moved between two even-larger energy minima at this point in the trajectory. A longer simulation would give more insight into these larger conformational ensembles, and whether this shift is simply a move away from the crystal conformation towards that found in aqueous solution. Therefore, convergence must be operationally defined in the context of the length of the simulation. Absolute convergence to a global energy minimum is impossible to assess when the time scale of the simulation is comparable to or less than the conformational recursion time. This is simply the multiple minima problem restated in the context of molecular simulations.

Implications for Protein Structure Prediction

The implications of these data for modeling proteins with unknown structures are not surprising. It is difficult to predict the tertiary structure of protein that is not homologous to any other protein of known structure. A predicted model within an rms deviation of 3–4 Å would be considered very good by current standards. What would be the necessary length of a molecular dynamics simulation for a protein to find its "correct" structure 3 Å rms deviation away? We can place a speculative estimate of a lower bound on this by examining the rms deviation between 13,315 pairs of bpti structures as a function of time. This is illustrated in Figure 9. The choice of a log scale for the time dimension is arbitrary. By extrapolating from the range of deviations seen in our simulations, microseconds of simulation would be the extreme lower bound for a protein to move to a conformational ensemble 3 Å rms deviation away from its original position. This is both a computationally and physically reasonable estimate of the amount of time it would take an actual protein to move from a

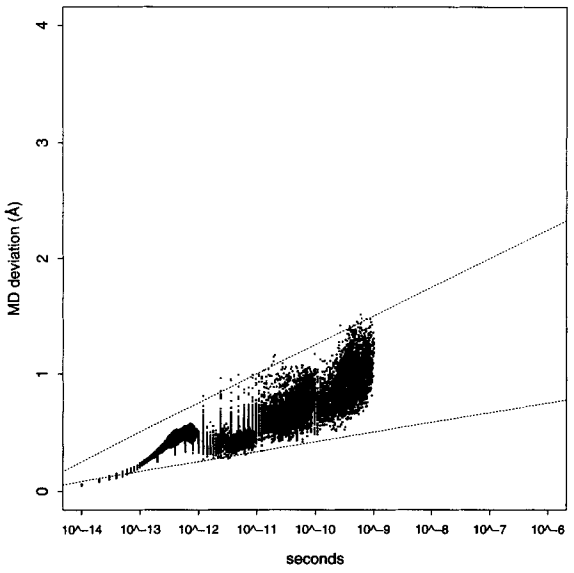


Fig. 9. C_α rms deviation values for pairs of bpti structures as a function of the log of their difference in time along the MD trajectory. The general trend is extrapolated with dotted lines.

partially misfolded state, but remains two to three orders of magnitude beyond currently accessible simulation lengths.

Constructed models of protein tertiary structure often move quite substantially on energy minimization and in the initial stages of dynamics, as high energy degrees of freedom are relaxed. Such short simulations may be useful in making the model more "protein-like," and may in some cases move the model closer to the correct structure. But such large changes are typically contained within the first tens or hundreds of picoseconds of simulation. Further simulation on this rugged landscape is not likely to improve the quality of the model. Even if we do not consider the inaccuracies of current potential functions, we must utilize other strategies to overcome the limits of conformational sampling imposed by current computational capabilities.

CONCLUSION

The energy minimization of structures drawn from the trajectory has given us some insight into the sizes of features of the conformational landscape of bpti, in particular the small features explored at a 1 ps time scale. For larger features explored at longer time scales, energy minimization does not change the overall interpretation of substates and their boundaries. The large amount of computer time required to bring the protein structures to a very low energy gradient limits the generality of energy minimization as a tool for analyzing long MD trajectories. However, it is clear that minimization does enhance the definition between substates.

Clustering is a useful method of characterizing

molecular conformations and identifying conformational substates. It can guide the analysis to focus on the most important changes seen in the course of the simulation. Simple matrices of rms deviation give insight into the overall extent of conformational sampling during a simulation, and they also emphasize the presence of substates and the transitions between them over the course of the calculation. The use of rms deviation matrices in this study clearly accentuates global changes in the protein structure at both small and large scales. These changes would be difficult to elucidate following a single degree of freedom, such as a dihedral angle or hydrogen bond.

The bpti molecule visits only 7 substates that vary predominantly in the rigid twist of a loop, which rotates 30° and does not move back to its original position within the time frame of the simulation. Each substate has a rms deviation of 0.7–1.0 Å from the others. In this MD trajectory, although the potential energy of the system has long since stabilized into an "equilibrium" value, conformational convergence has not been achieved. Simulations of large systems such as proteins must be evaluated carefully if complete conformational sampling of even a portion of the protein is desired.

ACKNOWLEDGMENTS

We would like to thank Jeanmarie Guenot and Peter Kollman for graciously providing the coordinates of the bpti trajectory analyzed in this paper. Alexis Falicov, Tom Cheatham, David Ferguson, and Roland Dunbrack provided very helpful comments. This work was supported by a grant from the National Institutes of Health and an AASERT training award from the Advanced Research Projects Agency of the Department of Defense.

REFERENCES

1. Frauenfelder, H., Sligar, S.G., Wolynes, P.G. The energy landscapes and motions of proteins. Science 254:1598–1603, 1991.
2. Hong, M.K., Braunstein, D., Cowen, B.R., Frauenfelder, H., Iben, I.E.T., Mourant, J.R., Ormos, P. Scholl, R., Schulte, A., Steinbach, P.J., Xie, A., Young, R.D. Biophys. J. 58:429–436, 1990.
3. Gerstein, M., Lesk, A., Chothia, C. Structural mechanisms for domain movements in proteins. Biochemistry 33:6739–6749, 1994.
4. Elber, R., Karplus, M. A molecular dynamics analysis of myoglobin. Science 235:318–321, 1987.
5. Kuriyan, J., Osapay, K., Burley, S.K., Brunger, A.T., Hendrickson, W.A., Karplus, M. Exploration of disorder in protein structures by X-ray restrained molecular dynamics. Proteins 10:340–58, 1991.
6. Guenot, J. Molecular dynamics and electrostatic modeling of proteins: Methodologies and applications for conformational sampling, structure prediction, structure refinement and macromolecular association. University of California, San Francisco, 1993.
7. Guenot, J., Kollman, P.A. Conformational and energetic effects of truncating nonbonded interactions in an aqueous protein dynamics simulation. J. Comp. Chem. 14:295–311, 1993.
8. Gordon, H.L., Somorjai, R.L. Fuzzy cluster analysis of molecular dynamics trajectories. Proteins 14:249–264, 1992.
9. Karpen, M.E., Tobias, D.J., Brooks, C.L. III. Statistical clustering techniques for the analysis of long molecular dynamics trajectories. Biochemistry 32:412–420, 1993.
10. Murray-Rust, P., Rafferty, J. Computer analysis of molecular geometry, Part VI: Classification of differences in conformation. J. Mol. Graphics 3:50–59, 1985.
11. Perkins, T.D.J., Barlow, D.J. RAMBLE: A conformational search program. J. Mol. Graphics 8:156–162, 1990.
12. Allen, F.H., Doyle, M.J., Taylor, R. Automated conformational analysis from crystallographic data. 3. Three-dimensional pattern recognition within the Cambridge Structural Database system: implementation and practical examples. Acta Cryst. B47:50–61, 1991.
13. Perkins, T.D.J., Dean, P.M. An exploration of a novel strategy for superposing several flexible molecules. J. Comput.-Aided Mol. Design 7:155–172, 1993.
14. Laaksonen, L. A graphics program for the analysis and display of molecular dynamics trajectories. J. Mol. Graphics 10:33–34, 1992.
15. Shenkin, P.S., McDonald, D.Q. Cluster analysis of molecular conformations. J. Comput. Chem. 8:899–916, 1994.
16. Cohen, F.E., Sternberg, M.J.E. On the prediction of protein structure: The significance of the root-mean-square deviation. J. Mol. Biol. 138:321–333, 1980.
17. Levitt, M. Molecular dynamics of native protein II. Analysis and nature of motion. J. Mol. Biol. 168:621–657, 1983.
18. Amadei, A., Linssen, A.B.M., Berendsen, H.J.C. Essential dynamics of proteins. Proteins 17:412–425, 1994.
19. Hayward, S., Kitao, A., Go, N. Harmonic and anharmonic aspects in the dynamics of BPTI: A normal mode analysis and principal component analysis. Protein Sci. 3:936–943, 1994.
20. Wlodawer, A., Deisenhofer, J., Huber, R. Structure of form III crystals of bovine pancreatic trypsin inhibitor. J. Mol. Biol. 193:145–156, 1987.
21. Guenot, J., Kollman, P.A. Molecular dynamics studies of a DNA-binding protein: 2. An evaluation of implicit and explicit solvent models for the molecular dynamics simulation of the *Escherichia-coli trp* repressor. Protein Sci. 1:1185–1205, 1992.
22. Pearlman, D.A., Case, D.A., Caldwell, J.C., Seibel, G.L., Singh, U.C., Weiner, P., Kollman, P.A. AMBER 4.0, University of California, San Francisco, 1991.
23. Arnold, G.E., Ornstein, R.L. An evaluation of implicit and explicit solvent model systems for the molecular dynamics simulation of bacteriophage T4 lysozyme. Proteins 18:19–33, 1994.
24. Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. Acta Cryst. A34:827–828, 1978.
25. Statistical Sciences, Inc. S-PLUS User's Manual, Version 3.0. Seattle: Statistical Sciences, Inc, 1992.
26. Davison, M.L. "Multidimensional Scaling." New York: Wiley, 1983.
27. Kossiakoff, A.A., Randal, M., Guenot, J., Eigenbrot, C. Variability of conformations at crystal contacts in BPTI represent true low-energy structures: Correspondence among lattice packing and molecular dynamics structures. Proteins 14:65–74, 1992.
28. Noguti, T., Go, N. Structural basis of hierarchical multiple substates of a protein. Proteins 5:97–138, 1989.
29. Levy, R.M., Sheridan, R.P., Keepers, J.W., Dubey, G.S., Swaminathan, S., Karplus, M. Molecular dynamics of myoglobin at 298 degrees K. Results from a 300-ps computer simulation. Biophys. J. 48:509–518, 1985.