Statistical Clustering Techniques for the Analysis of Long Molecular Dynamics Trajectories: Analysis of 2.2-ns Trajectories of YPGDV†

Mary E. Karpen, Douglas J. Tobias,‡ and Charles L. Brooks III*

Department of Chemistry, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

Received June 18, 1992

ABSTRACT: The microscopic interactions and mechanisms leading to nascent protein folding events are generally unknown. While such short time-scale events are difficult to study experimentally, molecular dynamics simulations of peptides can provide a useful model for studying events related to protein folding initiation. Recently, two extremely long molecular dynamics simulations (2.2 ns each) were carried out on the pentapeptide Tyr-Pro-Gly-Asp-Val [Tobias, D. J., Mertz, J. E., & Brooks, C. L., III (1991) *Biochemistry 30,* 6054–6058] that forms stable reverse turns in solution. Tobias et al. examined folding events in this large system (~30 000 conformations) using traditional methods of trajectory analysis. The shear magnitude of this problem prompted us to develop an automated approach, based on self-organizing neural nets, to extract the key features of the molecular dynamics trajectory. The neural net is used to perform conformational clustering, which reduces the complexity of a system while minimizing the loss of information. The conformations were grouped together using distances in dihedral angle space as a measure of conformational similarity. The resulting clusters represent "conformational states", and transitions between these states were examined to identify mechanisms of conformational change. Many conformational changes involved the rotation of only a single dihedral angle, but concerted angle changes were also found. Most of the conformational information in the 30 000 samples from the full trajectories was retained in the relatively few resultant clusters, providing a powerful tool for analysis of an expanding base of large molecular simulations.

As computer technologies advance, increasingly longer molecular dynamics and Monte Carlo simulations are achievable. Currently, the longest molecular dynamics simulations have extended into the nanosecond range (Soman et al., 1991; Tobias et al., 1991a). It is becoming essential to develop methods which reduce the complexity of data resulting from these long simulations while retaining the relevant information. Although techniques for analyzing dynamical motion within dynamics trajectories are well developed (e.g., correlation analysis, Fourier transforms), self-similar conformations within a trajectory have traditionally been identified by time-consuming frame-to-frame graphics analysis. Our goal is to develop an automated means of identifying recurring conformations within dynamics trajectories, which can subsequently form a framework for analyzing stabilizing interactions and conformational transitions.

Statistical clustering techniques are well suited to objectively organize trajectory data. These methods classify initially unclassified data and have been used in many fields to discover the underlying structure in complex data. Numerical taxonomy was the first field to widely use clustering, constructing binary trees to describe evolutionary relationships between species (Everitt, 1980). Nonhierarchical clustering methods that find a single, optimal partition of data based on a set of target criteria (e.g., number of clusters, cluster size) have also been developed (Duda & Hart, 1973; Everitt, 1980). These methods have an advantage over many hierarchical techniques in that data can be optimally reassigned after initial assignment to a cluster.

Clustering techniques have been previously used in the analysis of conformational data, particularly in identifying recurring conformations in folded protein structures, obtained from X-ray diffraction data (Karpen, 1992; Rooman et al., 1990; Unger et al., 1989). Nonoptimized classification methods have also been used to group conformational data. For example, classes of minimum energy configurations obtained in a conformational search of helices have been extracted (Polinsky et al., 1992). Classification of dynamics trajectory data has been carried out by McKelvey et al. (1991), who grouped conformations of a peptide by assigning each residue to a region of the ϕ–ψ map, following earlier ideas of Zimmerman et al. (1977). Levitt (1983) grouped conformations from a molecular dynamics trajectory of the protein PTI, after projecting the conformations onto a 2-D surface. These analyses have motivated our current use of optimized clustering to analyze long trajectories from molecular simulations, where "redundant" conformations in a trajectory are grouped together, thereby allowing us to examine transitions common to many dynamics frames.

As a prototype system, we explored clustering algorithms applied to data from a long molecular dynamics simulation by Tobias et al. (1991a) of the pentapeptide Tyr-Pro-Gly-Asp-Val (YPGDV). Some experimental evidence suggests that in globular proteins secondary structure forms early during the folding process (at less than millisecond time scales) and segments of secondary structure or microfolding domains then coalesce to form tertiary structure (Dobson, 1991; Kim & Baldwin, 1990, and references therein). Since initial protein folding is presumably a local event, peptides can be used as a simple model system for studying secondary structure formation. NMR studies have shown that the YPGDV pentapeptide forms relatively stable reverse turns in solution, with approximately 50% of the population in a type II turn (Dyson et al., 1988). To identify conformational transitions

† This work was supported by NIH Grants GM37554 (C.L.B.) and GM14525 (M.E.K.), a Molecular Simulations Inc. Fellowship (M.E.K.), and an NIH predoctoral training grant and an internship at Cray Research Inc. (D.J.T.). C.L.B. is an A. P. Sloan Foundation Fellow (1990–1993).

‡ Present address: Department of Chemistry, University of Pennsylvania, Philadelphia, PA 19104

0006-2960/93/0432-412$04.00/0

potentially important in the initiation of protein folding, as well as the time scales of these events, Tobias et al. (1991a) investigated the dynamical behavior of YPGDV via molecular dynamics in explicit solvent. Two 2.2-ns trajectories were carried out, one with the peptide initially in a type II turn (the "turn" trajectory) and a second with the peptide initially in an extended conformation (the "extd" trajectory). Conformations were stored every 0.15 ps, and reverse turn formation and dissolution were observed (Tobias et al., 1991a).

Close to 15 000 conformations were saved in each of the two simulations, and several conformational transitions occurred. Thus, this data set is a good test case for automated analysis via clustering techniques. In this paper, we present results from clustering both the turn and extd trajectories of Tobias et al. (1991a), using a nonhierarchical clustering method, ART-2′ (Carpenter & Grossberg, 1987; Pao, 1989), to cluster conformations in the full parameter space of dihedral angles. This method is stepwise optimal and locates clusters in conformational space so as to minimize the distance between the conformations within a cluster and the cluster center, i.e., the average cluster conformation. We examined the cluster centers to identify interactions which stabilize peptide conformation, and mechanisms of turn unfolding and folding were inferred from transitions between the clusters. As a visualization tool, we projected the clusters onto a 2-D surface, using multidimensional scaling. The utility of the clustering method was assessed by comparing our results from cluster analysis to those obtained by Tobias et al. (1991a) using traditional analyses of dynamics trajectories.

METHODS

Molecular Dynamics. Molecular dynamics (MD) simulation methods were used to generate the peptide conformations we analyzed. We briefly describe the MD simulation protocol; details can be found in Tobias et al. (1991a). The peptide (YPGDV) was initially built in an ideal type II turn and equilibrated in a dielectric continuum ($\epsilon = 50$) for 150 ps (all simulations were performed at 300 K). It was then solvated in a 29-Å box of TIP3P water molecules (Jorgensen et al., 1983). Using periodic boundary conditions, an MD simulation was carried out for 2.2 ns, yielding the turn trajectory. The extd trajectory used the same protocol, except the peptide started in an extended conformation, with backbone dihedral angles equal to those of an ideal parallel β strand. The peptides were built with charged N- and C-terminal groups (NH_3^+ and CO_2^-, respectively) in accordance with the peptide used in the experiments of Dyson et al. (1988).

Clustering Algorithm. ART-2′ is a nonhierarchical clustering algorithm based on a self-organizing neural net (Carpenter & Grossberg, 1987; Pao, 1989). We chose this algorithm to cluster the conformations by structural similarity rather than a hierarchical method both because it produces a single, optimized partition of the trajectory data and because much less memory is required in its implementation. ART-2′ optimizes cluster assignment subject to a constraint on cluster radius, such that no member of a cluster is more than a specified distance from the cluster center. This optimization is carried out as an iterative minimization procedure that minimizes the mean-square error between the cluster center and the cluster members. Alternatives to the radius constraint, such as constraints on the number of clusters, are found in other clustering algorithms (Duda & Hart, 1973); we have chosen cluster radius because the number of clusters is not known a priori and because we expect the natural clusters to be of similar extent in conformational space.

We now describe the ART-2′ algorithm in detail. A conformation, j, is described as a vector of N parameters, $\mathbf{x}_j = [x_{1j}, ..., x_{Nj}]$, for example, dihedral angles or atomic coordinates. Cluster k, C_k, is characterized by its *cluster center*, $\mathbf{c}_k = [c_{1k}, ..., c_{Nk}]$, which is the average conformation of its M members, that is

$$\mathbf{c}_k = \frac{1}{M}\sum_{\mathbf{x}_j \in C_k} \mathbf{x}_j \tag{1}$$

If the $\mathbf{x}_j$ contain angle data, the summation is adjusted to take into account angle periodicity. Conformation j is compared to cluster center k by the Euclidean distance d_{jk}

$$d_{jk} = [(\mathbf{x}_j - \mathbf{c}_k)^T(\mathbf{x}_j - \mathbf{c}_k)]^{1/2} \tag{2}$$

where $(...)^T$ denotes a vector transpose.

The criteria for clustering is that all conformations in a cluster must be within a specified threshold distance, or *cutoff radius*, from the cluster center. The algorithm compares each conformation $\mathbf{x}_j$ to the set of all cluster centers, using eq 2, and the cluster C_k with the minimum d_{jk} (i.e., the closest cluster) is determined. If the distance from the conformation to cluster C_k is within the cutoff radius, the conformation is assigned to that cluster. When the distance exceeds the threshold, the conformation forms a new cluster. In the initial "learning" phase, cluster centers are recalculated as each new member is added. In subsequent "refining" phases, cluster centers are not updated until all conformations are read in and assigned. The clustering is complete when cluster membership does not change between iterations.

The ART-2′ algorithm is susceptible to convergence to a local minimum, dependent on the initial order of the conformations to be clustered. In our calculations, the conformations were presented in the order they appeared in the dynamics trajectory. To test the effect of this ordering, we clustered conformations initially in a random order; a similar set of clusters resulted, showing that the clustering found stable groupings.

The clustering algorithm requires a choice of both a set of parameters to describe an object, in this case peptide conformation, and a cutoff radius. We chose to cluster conformations using backbone and side-chain dihedral angles. Rotations about dihedral angles are the soft degrees of freedom in protein folding, so this measure is well suited to identify transition events. In addition, clustering based on differences in internal coordinates, such as dihedral angles, eliminates the need to superpose structures into a common reference frame, required in Cartesian-based clustering methods. One disadvantage to the use of this metric is related to the fact that small dihedral angle changes near the center of geometry of a structure can lead to relatively large changes in conformation. We discuss issues which arise due to this complication below.

The cutoff radius was chosen so as to separate conformations that differed by large transitions in dihedral angle, but group together those that differed by comparatively small thermal fluctuations. Since thermal fluctuations in backbone and side-chain dihedral angles of proteins are on the order of 15° (Brooks & Karplus, 1989), differences in conformation due to thermal fluctuations alone are expected to be at least 42° when clustering the eight backbone dihedral angles (ψ_1, ϕ_2, ψ_2, ϕ_3, ψ_3, ϕ_4, ψ_4, ϕ_5, indexed by residue position, Tyr1–Val5). We chose a cutoff radius of 170°, well above the thermal noise level, thereby requiring a large dihedral angle transition (~120°) or two or more smaller transitions for conformation separation. As will be shown, most conformations within a cluster are within the thermal noise; a smaller radius caused the "transition" conformations, those on the path from one conformational class to another, to cluster separately, greatly increasing the number of clusters, albeit with low populations.

To cluster the 13 backbone and side-chain dihedral angles, a similar radius of 162° was used, which averages to a 45° difference per angle; differences of 40° and 50° per angle were also used to cluster the turn trajectory, but these cutoffs resulted in a relatively large number (25) and small number (6) of clusters, respectively. The side-chain dihedrals clustered were χ^1 and χ^2 of Tyr1, χ^1 and χ^2 of Asp4, and χ^1 of Val5; these angles cause the major changes in peptide conformation. The χ^2 angles of Tyr and Asp are essentially conformationally invariant under a 180° rotation and were shifted to the range of 0°–180° before clustering.

ART-2′ has been implemented in the molecular mechanics program CHARMM (Brooks et al., 1983). In addition to dihedral angles, a variety of parameters may be clustered, including internal coordinates and potential energies.

Multidimensional Scaling. It is useful to be able to visualize the relative location of clusters in dihedral space, but only low dimensional spaces are amenable to such visualization. Thus, we used multidimensional scaling to project the clusters from n-dimensional dihedral angle space onto a two-dimensional plane. This is similar to the scaling performed earlier by Levitt (1983) but has the advantage that we first cluster the conformations in full parameter space and then scale. The projection was done so as to minimize the error (Φ) in intercluster distances in the reduced dimensional space (d_{ij}^*) with respect to these distances in the full dihedral angle space (d_{ij})

$$\Phi = \sum_{i=1}^{N}\sum_{j=1}^{N}(d_{ij} - (\alpha + \beta d_{ij}^*))^2 \qquad (3)$$

where N is the number of clusters. The objective function Φ is minimized with respect to the coordinates of each cluster in 2-D space and the scale factors α and β (IMSL, 1987; Takane et al., 1977).

Analysis of Trajectories. To detect concerted dihedral angle changes, we calculated the correlation coefficient ρ for fluctuations in a pair of dihedral angles $\Delta\theta_1$, $\Delta\theta_2$ as

$$\rho(\Delta\theta_1,\Delta\theta_2) = \frac{\langle \Delta\theta_1 \Delta\theta_2 \rangle}{\langle \Delta\theta_1^2 \rangle^{1/2} \langle \Delta\theta_2^2 \rangle^{1/2}} \qquad (4)$$

where $\langle ... \rangle$ denotes the expectation value.

RESULTS AND DISCUSSION

Dihedral Angle Time Series. The dihedral angle time series used to cluster the conformations are given in Figure 1. Several large transitions in both trajectories are apparent; smaller fluctuations about the mean, due to thermal noise, are evident as well. The central residue, Gly3, most frequently underwent large backbone dihedral angle changes (Figure 1a,b). This is expected from the greater conformational freedom of this amino acid due to its lack of a side chain. Conversely, the ϕ angle of Pro2 had only minor variations, because of constraints by the pyrrolidine ring. Several side-chain conformational transitions also occurred (Figure 1c,d). In general, the pattern of dihedral angle changes is complex, and the manual analysis of these time series for self-similar conformations and transitions is extremely tedious.

Anatomy of a Turn Unfolding. Tobias et al. (1991a) observed that the type II turn unfolded during the turn trajectory. In what follows, we use the clusters computed from our analysis of this trajectory as a framework for describing the unfolding process. We clustered the trajectory for two reasons: (i) to see if information about the turn trajectory would be preserved in the clusters and (ii) to identify transitions along the unfolding pathway. To illustrate our observations, we give a hierarchical description of the peptide conformation, first using only backbone dihedral angles and then using both backbone and side-chain dihedral angles. Clustering based on parameters relating to solvent structure would be a further step in the hierarchical description, though in this paper we focus only on the peptide conformation.

An overview of the clusters formed is given in Table I. Six clusters, T1–T6, resulted from classifying the turn trajectory by ϕ and ψ backbone dihedral angles. The number of conformations within each cluster, which varied from 5080 (T1) to 1066 (T6), could in principle provide thermodynamic information on conformational preferences. In the 2-ns simulation, however, the system cannot be considered to have reached equilibrium, and thus free energies could not be directly extracted from the population statistics. The standard deviations are substantially smaller (51°–68°) than expected from a cluster of radius 170° with uniformly distributed members (120°), indicating fairly compact clusters.

The nature of the recurring conformations within the compact clusters is indicated by the average cluster conformation, built from the clusters' average dihedral angles using ideal bond lengths and angles and a planar peptide bond. These cluster conformations are given in Figure 2 in the order of their appearance in the trajectory. The first three clusters, T1, T2, and T6, which occupied the first two-thirds of the trajectory, are all type II turns. The last cluster to occur, T5, is completely unfolded and forms a polyproline helix, commonly seen in extended structures within a protein that are not in a β sheet (Adzhubei et al., 1987; Karpen, 1991; Richardson & Richardson, 1989). The transitional clusters, T3 and T4, represent partially unfolded conformations which provide a bridge between the folded and unfolded conformations. Thus the cluster conformations demonstrate a turn unfolding event.

A more detailed picture of cluster evolution as a function of time is given in Figure 3. The first conformation of the trajectory was assigned to T2, and multiple transitions, or "toggling", between T1 and T2 took place for the first 900 ps. The peptide then shifted to toggling between T6 and T3 (900–1400 ps) and finally to conformations T4 (1400–1650 ps) and T5 (1650–2200 ps). Substantial toggling occurred for most of these transitions, with the exception of the T2 to T6 transition at 900 ps.

We examined the cluster conformations in more detail to identify interactions stabilizing the backbone conformation[1] and compared these observations to those of Tobias et al. (1991a), shown in the time line of Figure 3. The initial cluster, T2, is a type II turn with the turn hydrogen bond between the backbone carbonyl oxygen of residue 1 (O_1) and the amide proton of residue 4 (H_4) (Figure 2). In this conformation, the positive amino terminus and the negative carboxyl terminus are about 3 Å apart, forming a salt bridge. The salt bridge lengthened to greater than 4 Å in the next cluster, T1, which is also a type II turn but with the C-terminus now interacting with the amide proton of Gly (H_3) at the top of the turn. In addition to the turn hydrogen bond, T1 also had a hydrogen bond between O_1 and H_5, further stabilizing the type II turn (Figure 2). The T6 cluster was the last type II turn in the trajectory, which differed from T1 and T2 in that residue 4 had a β-type conformation rather than an α-type conformation, which extends the C-terminus away from the turn (Figure 2). Thus, this conformational transition is the first step in the unfolding of the peptide.

[1] We use "stability" to refer to kinetic stability; inference about thermodynamic stability cannot be made given the relatively short simulation time.

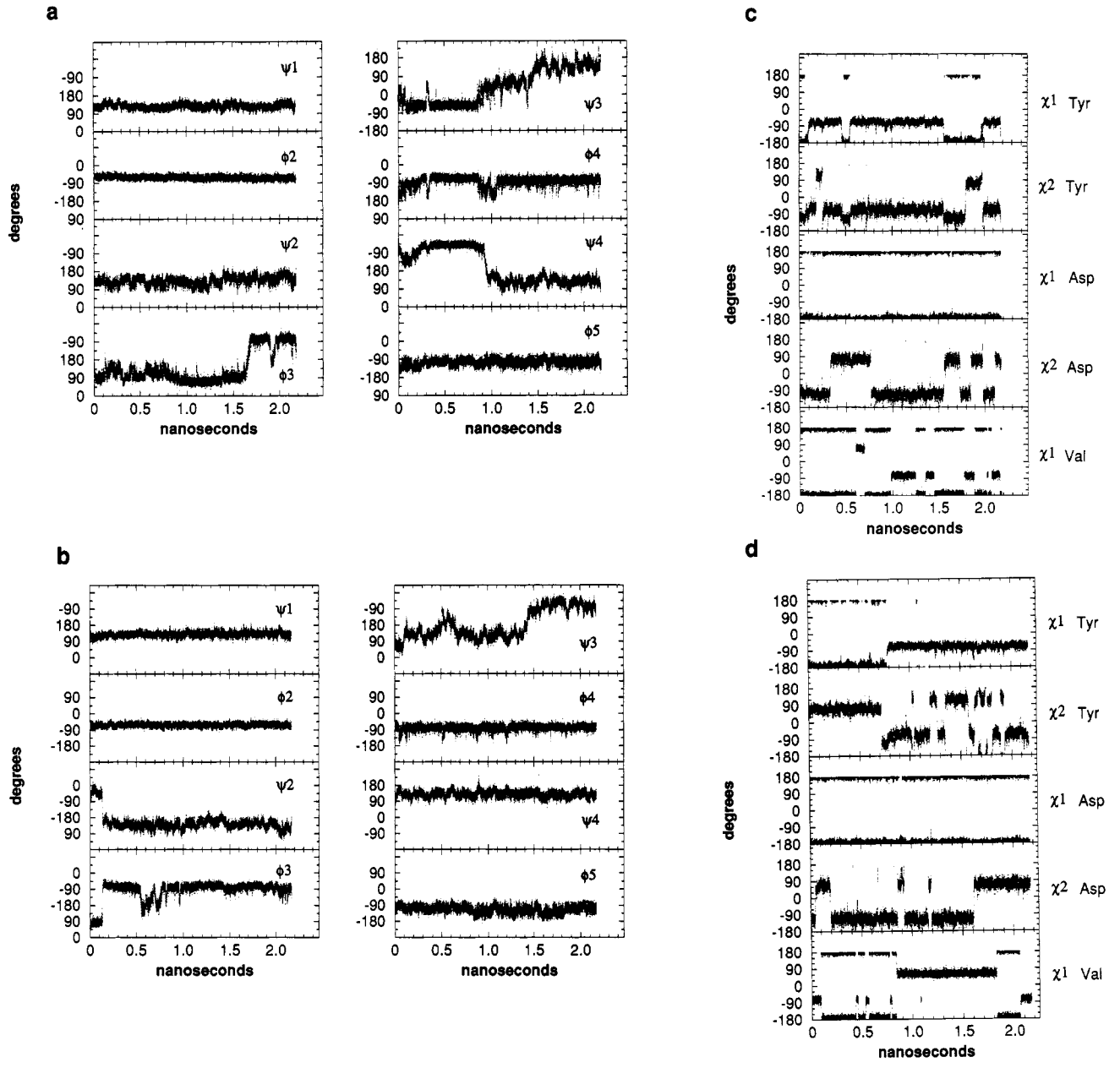

FIGURE 1: Dihedral angles as a function of time: backbone ϕ and ψ angles from the (a) turn and (b) extd trajectories; side-chain dihedral angles from the (c) turn and (d) extd trajectories.

Table I: Clusters Resulting from Clustering Backbone Dihedral Angles of the Turn Trajectory

			average dihedral angles								
			Tyr	Pro		Gly		Asp		Val	
cluster	no. of members	SD[a]	ψ_1	ϕ_2	ψ_2	ϕ_3	ψ_3	ϕ_4	ψ_4	ϕ_5	structure type
T1	5080	54	126	−59	127	115	−55	−73	−61	−108	type II turn
T2	1147	59	133	−59	131	86	7	−133	−97	−113	type II turn
T3	2501	51	128	−62	122	76	63	−82	120	−106	
T4	1504	65	129	−64	140	118	145	−86	137	−101	
T5	3231	57	127	−65	144	−85	137	−82	127	−105	β strand
T6	1066	68	131	−61	130	83	22	−118	146	−98	type II turn

[a] Standard deviation of distances from the cluster center to each conformation assigned to the cluster.

T3 and T4 were partially unfolded conformations, with no direct backbone–backbone hydrogen bonds. T3 had an average conformation similar to T6, but with the turn hydrogen bond replaced by a solvent-separated hydrogen bond. In T4, the backbone of Gly became more extended, and the O_3 atom had replaced the H_4 atom in the solvent-separated bond with O_1. Finally, T5 had an extended conformation equivalent to a polyproline helix.

When this sequence of events observed from the analysis of cluster evolution is compared with those observed by Tobias et al. (Figure 3), a high degree of correlation is seen. Tobias et al. noted a shift in C-terminus orientation at ~250 ps, which corresponds to the T2 to T1 transition. They found that the turn "dissolved" at 1350 ps, which corresponds with the disappearance of T6. They also observed "U-shaped" structures starting at 1490 ps (T4 conformation) and unfolded

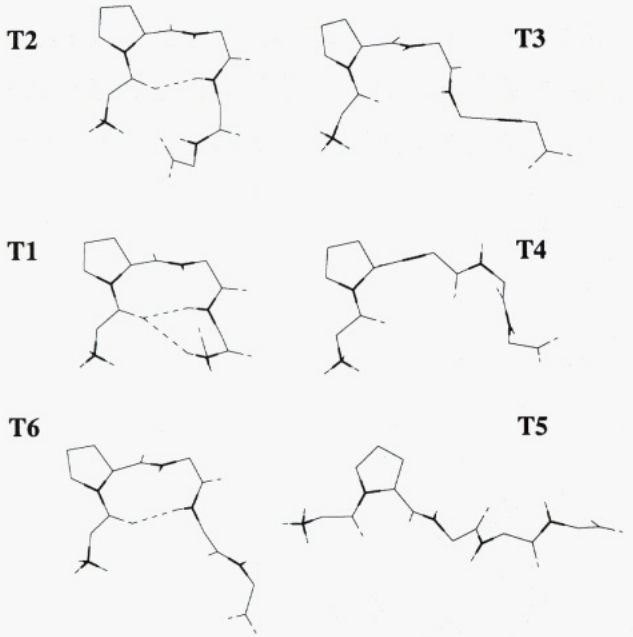

FIGURE 2: Average conformations of the turn trajectory backbone clusters. Backbone atoms are shown as well as the side-chain atoms of Pro2. Nitrogen is denoted by thick lines, and dashed lines in T2, T1, and T6 denote hydrogen bonds.

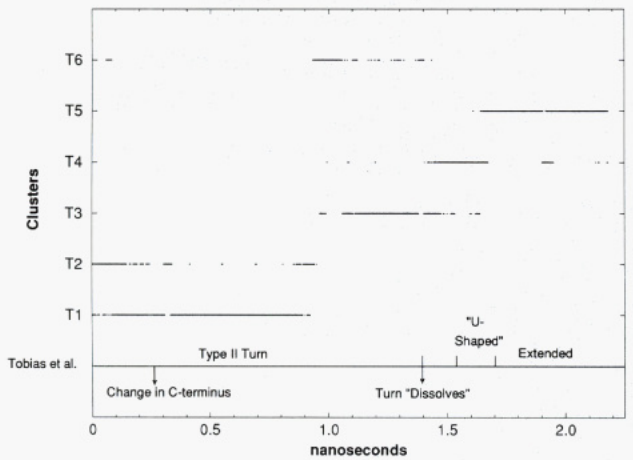

FIGURE 3: Cluster assignments as a function of time for the turn trajectory (backbone clusters). The cluster assigned to the conformation at each time point is indicated. The Tobias et al. time line gives the observations of Tobias et al. (1991a).

conformations after 1650 ps (T5 conformation). Hence, the major unfolding events of the turn trajectory are preserved in the cluster structure.

The transitions that occurred between the clusters point to possible mechanisms for turn unfolding and are examined below. To better visualize the relationship between the clusters, we used multidimensional scaling (see Methods) to project the cluster centers onto a two-dimensional surface (Figure 4). The projection was optimal in the sense of preserving intercluster distances. Some loss of information is inherent in such a representation; nonetheless, many features of cluster space are readily apparent.

The distances between the clusters in Figure 4 best reflect, in a least squares sense, the true distances in multidimensional dihedral angle space. The radius about each cluster center approximately equals the conformational standard deviation of members within the cluster. The thickness of lines connecting the clusters is proportional to the number of transitions between each cluster pair, and each transition is labeled with the dihedral angles that underwent major changes

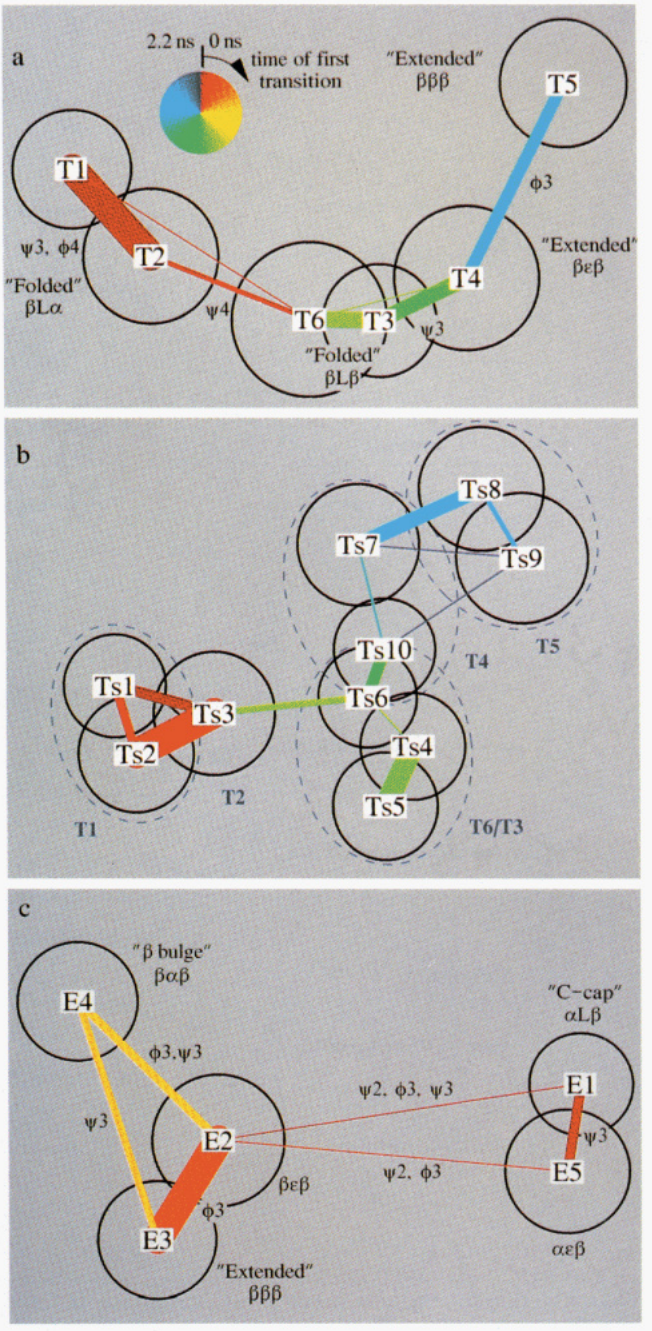

FIGURE 4: Two-dimensional projection of cluster space by multidimensional scaling. The label of each cluster is centered at the cluster's location in two-dimensional space, and the relative number of transitions between two clusters is indicated by the thickness of the line connecting them (no line denotes no transitions). The time of the first transition between two clusters is encoded as a color, determined from the color wheel in panel a. Each transition is labeled by the dihedral angles that undergo the major change. The radius of each circle is approximately equal to the standard deviation of conformations from the cluster center (as given in Tables I and II), projected into 2-D space by scaling each standard deviation by the ratio of the smallest intercluster distance in 2-D space (e.g., the T3–T6 distance in panel a) to this same distance in full dihedral angle space. (a) Projection of backbone clusters from the turn trajectory. Clusters are labeled with their general conformational state (folded or extended) and the conformation of the center three residues (Pro, Gly, and Asp). (b) Projection of backbone plus side-chain clusters from the turn trajectory. The dashed ellipses enclose those clusters that are subsets of the backbone turn clusters, as labeled. (c) Projection of backbone clusters from the extd trajectory.

in the cluster transition. The general order of transitions with respect to time is indicated by color (note that our coloring of "transition tie lines" is based on first-passage time).

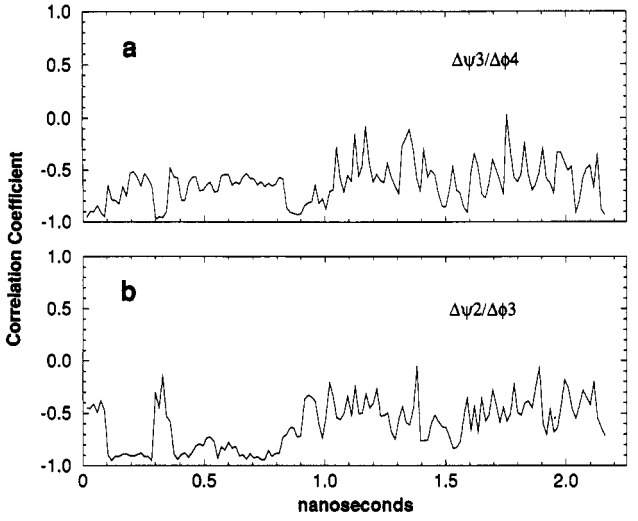

FIGURE 5: Correlation coefficient for a pair of dihedral angle fluctuations, calculated over a 30-ps sliding window: (a) fluctuations in ψ_3 vs ϕ_4; (b) fluctuations in ψ_2 vs ϕ_3.

The number of transitions between a pair of clusters correlates with the distance between them (Figure 4). The greatest number of transitions in the turn trajectory occurred between T1 and T2 (299), followed by transitions T6/T3 (163), T3/T4 (157), T4/T5 (123), and T2/T6 (40) (Figure 4a). Only one transition occurred between T1/T6 and between T6/T4.

Most transitions involved a change in a single dihedral angle. The transition between clusters T1 and T2 is an exception to this, as changes in both ψ_3 and ϕ_4 occurred. These angles flank the peptide bond between Gly3 and Asp4, and their concerted, anticorrelated change rotates the peptide group, leaving the neighboring C_α atoms in approximately the same position. This peptide "flip" is a common feature in protein dynamics (Levitt, 1983) and is readily apparent in Figure 5a, which gives the correlation coefficients for $\Delta\psi_3$ vs $\Delta\phi_4$ calculated over 30-ps windows. In the regions where T2/T1 transitions occurred (0–100, 300–350, and 900–950 ps), the anticorrelation between $\Delta\psi_3$ and $\Delta\phi_4$ increased (Figure 5a), with the correlation coefficient changing from about –0.6 to –0.9. In these same regions, $\Delta\psi_2$ and $\Delta\phi_3$ (i.e., rotations of the Pro–Gly peptide bond) became less correlated (Figure 5b) due to small changes in ϕ_3 independent of ψ_2, probably required for optimal hydrogen bonding. Note that, at times greater than about 1 ns, fluctuations in the correlation coefficient increased for both sets of angles. This reflects the fact that the C_α positions were no longer constrained by the turn hydrogen bond, allowing ψ_i to move independently of ϕ_{i+1}. Thus, concerted dihedral angle fluctuations are a function of peptide conformation.

In contrast to the T1/T2 transition, which is difficult to distinguish from thermal noise (Figure 1a), the transition from T2 to T6 involves a very large change in ψ_4, from about –100° to 150°. This changes Asp4 from an approximate α conformation to a β conformation, initiating peptide unfolding. As can be seen in Figure 1a, the transition goes through –180°/180° rather than the "bridge" region between α and β at 0°. This is somewhat surprising since, for non-glycine amino acids, the bridge region is of lower energy than the –180°/180° region (Hermans et al., 1990) and is more frequently occupied in proteins (Richardson & Richardson, 1989). Indeed, dynamics studies of the unfolding of α helices and type I turns have α to β transitions traversing the bridge region (Czerminski & Elber, 1989; DiCapua et al., 1990; Lazaridis et al., 1991). From graphics analysis of the type II turn conformation, it appears the rotation of ψ_4 through 0° requires a close approach between the peptide carbonyl and the carboxyl side chain of Asp4. This unfavorable electrostatic interaction may raise the free energy barrier to rotations through the bridge region. Thus, the side chain at residue 4 may influence the folding pathway of the peptide.

Although T1 and T2 had similar conformations, it was the T2 conformation that interchanged with T6 (40 T2/T6 transitions vs 1 T1/T6 transition). The T1 conformation, with its C-terminal interaction with the Gly amide proton and its stronger O_1–N_5 hydrogen bond, would presumably stabilize the type II turn against a transition to the T6 region by constraining ψ_4. A T2 conformation may be an obligatory step between a T1 to T6 transition. One transition between T1 and T6 did occur (at 62 ps), though it is possible that a T2 conformation occurred between the conformations but was not stored.

The transition between T6 and T3 had small adjustments in three dihedral angles, ψ_3, ϕ_4, and ψ_4. Frequent transitions occurred between these two clusters, at a rate of about 300 transitions/ns when normalized by the time spent in the cluster pair. This suggests a susceptibility of the T6 conformation to solvent insertion.

The transition between T3 and T4 involved changes in both ϕ_3 and ψ_3 of Gly, though no strong correlation in their transitions was found (data not shown). The backbone conformation of Gly changed from its conformation in the type II turn (denoted L due to its similarity to conformations in a left-handed α helix) to an ϵ conformation, an extended conformation energetically favorable only for glycine, causing further unfolding of the peptide. The smaller change in ϕ_3 preceded the larger change in ψ_3, perhaps breaking the solvent-separated O_1–H_4 hydrogen bond, allowing greater rotational freedom for ψ_3. For the T4/T5 transition, ϕ_3 undergoes a major change at about 1.6 ns, changing the Gly conformation to β. The peptide is completely unfolded at the end of the simulation.

Although the system is not at equilibrium, the cluster transitions can give qualitative information about the underlying free energy surface within the conformations visited. Two clusters with frequent transitions may correspond to conformations that occupy the same free energy basin or that are separated by a low, thermally accessible barrier. This may be true, for example, of the T1/T2 pair or the T3/T6 pair, both of which had frequent transitions. Conversely, the T2/T6 transition was relatively abrupt; in the region around 900 ps only 21 transitions occurred. The T2 and T6 clusters most likely represent two separate minima on the free energy surface.

Turn Trajectory Side-Chain plus Backbone Clusters. Side-chain conformations can strongly influence peptide conformational stability. To examine this effect, we clustered both backbone and side-chain dihedral angles, using a cutoff radius of 162°. Ten clusters resulted, Ts1–Ts10. These clusters are illustrated by the projection of 13-dimensional dihedral space onto a 2-D surface in Figure 4b. The clusters were generally subsets of the backbone clusters, as indicated on the figure.

One potentially important side-chain interaction apparent in the cluster conformations was the packing of Tyr1 against the Pro2 side chain. Ts1 and Ts2 had similar backbone conformations (that of T1), but they differed in Tyr1 side-chain conformation; in Ts1, the Tyr side chain packed against Pro, while in Ts2 it extended into the solvent. Other cluster pairs that had similar backbone conformation but differed in Tyr packing were Ts7, Ts10 (~T4), and Ts8, Ts9 (~T5). Valine underwent conformational changes in clusters corre-

Table II: Clusters Resulting from Clustering Backbone Dihedral Angles of the Extd Trajectory

			average dihedral angles								
			Tyr	Pro		Gly		Asp		Val	
cluster	no. of members	SD[a]	ψ_1	ϕ_2	ψ_2	ϕ_3	ψ_3	ϕ_4	ψ_4	ϕ_5	structure type
E1	630	48	116	–64	–40	81	61	–83	133	–98	$\alpha L\beta$
E2	1467	62	130	–62	143	–173	170	–78	134	–104	
E3	7053	57	129	–63	143	–77	126	–80	130	–115	β strand
E4	5036	57	132	–62	136	–84	–79	–78	127	–117	β bulge
E5	274	59	117	–64	–44	92	138	–110	135	–107	

[a] Standard deviation of distances from the cluster center to each conformation assigned to the cluster.

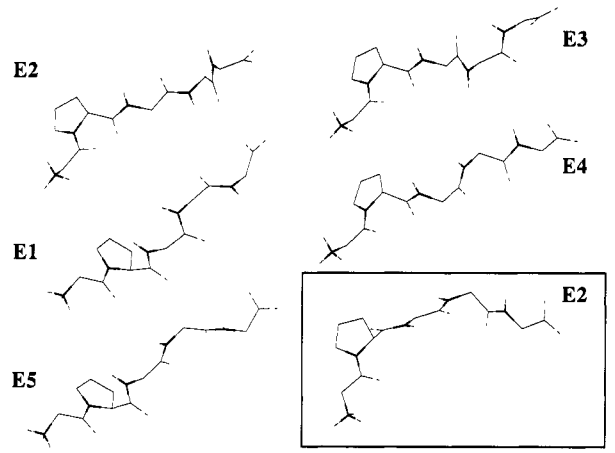

FIGURE 6: Average conformations of the extd trajectory backbone clusters. Thick lines denote nitrogen atoms. Inset: An different view of E2, which better shows its U-shaped conformation.

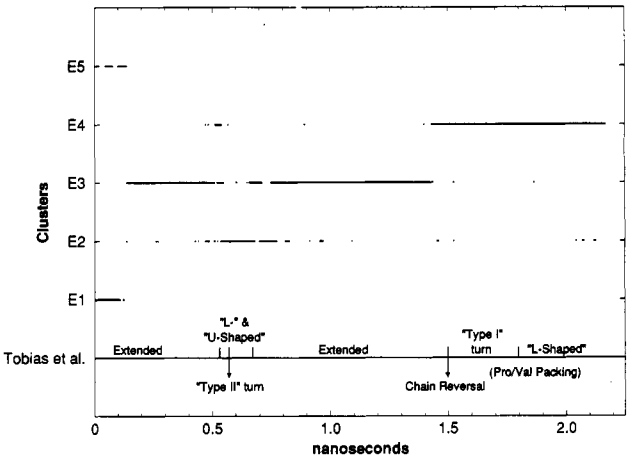

FIGURE 7: Cluster assignments as a function of time for the extd trajectory (backbone clusters). The cluster assigned to the conformation at each time point is indicated. The Tobias et al. time line gives the observations of Tobias et al. (1991a).

sponding to T3 and T6, though the change did not appear to substantially alter intrapeptide interactions. Finally, a solvent-separated salt bridge occurred between the negative Asp side chain and the positive N-terminus for the type II clusters (T1, T2, and T6) and for cluster T3, potentially stabilizing the folded conformation. There were no strong correlations between side-chain and backbone conformational changes, indicating these events were not tightly coupled.

To summarize the turn unfolding event, the pentapeptide started in a type II turn that was stabilized by interactions between the termini and the Asp side chain (T1 and T2). The backbone of the fourth residue then extended, disrupting termini/backbone interactions (T6). This conformation had a turn hydrogen bond that was more susceptible to attack by solvent, interchanging frequently with a similar conformation which had a solvent molecule inserted into the turn hydrogen bond. The turn hydrogen bond eventually broke, allowing the peptide to adopt a more extended family of conformations. The extended conformations evolved through solvent-separated intermediates, first with a solvent bridge between O_1 and H_4 (T3) and then with one between O_1 and O_3 (T4). Finally, the peptide adopted a fully extended polyproline conformation (T5).

Heterogeneity of Extended Structures in the Extd Trajectory. In order to observe folding as well as unfolding events, Tobias et al. (1991a) repeated the simulation, but with the pentapeptide initially in an extended conformation. We clustered this trajectory using backbone dihedral angles and the same radius as the turn trajectory clustering (170°). Five clusters resulted, E1–E5 (Table II), which had a wide range of populations, from 7053 members (E3) to 274 members (E5). Although the conformations were generally extended (Figure 6), several important differences emerged upon closer examination of the conformations.

E3 has a polyproline conformation, very similar to a β strand. E2 is conformationally similar to E3 but with the central glycine residue in the extended ϵ conformation. The remaining clusters have conformations that differ from a β strand by the orientation of their peptide groups. E1 has the conformation $\alpha L\beta$ (Figure 6). This conformation is often found at the C-termini of α helices, as it aligns three NH groups (NH_2, NH_3, and NH_4 here) which can hydrogen bond to the CO groups exposed at the helix C-terminus (Karpen, 1991; Schellman, 1980). In this case, however, residue 2 was a proline, which has no amide proton and therefore would not be expected to cap a helix. It is tempting to speculate, however, that similar conformations in a non-proline-containing sequence could spontaneously fold and initiate helix formation. E5 has a conformation similar to E1 but with Gly in an ϵ conformation rather than an L conformation.

In E4, the central Gly residue has an α conformation, forming a $\beta\alpha\beta$ conformation often found at β bulges (Karpen, 1991; Richardson et al., 1978). In this conformation the peptide groups are aligned rather than antialigned as in the β strand (note the relative position of NH and CO groups in E4 compared with E3 in Figure 6). Thus, several different extended conformations that commonly occur in proteins were adopted by the pentapeptide.

We now describe the transitions that occurred between the clusters from the extd trajectory. The cluster assignments as a function of time are given in Figure 7, and the 2-D projection of cluster space is given in Figure 4c. The general order of appearance of the clusters was E2–(E1/E5)–(E2/E3/E4). That is, E1/E5 formed a transition pair and E2/E3/E4 formed a "transition triplet" (Figure 4c). A single transition sent the peptide from E2 to E1 at the start of the trajectory, where the conformations toggled between E1 and E5 for about 120 ps before returning to E2 via E5. This is an example of a single transitional event from one well-populated region to another, returning again in a single step but by an alternate route. The two transitions required substantial changes in three dihedral

angles, a peptide flip involving an anticorrelated 180° rotation in both ψ_2 and ϕ_3 and a smaller change in ψ_3.

The Gly conformation was mainly responsible for all cluster transitions other than E2/E1 and E5/E2. In 2.2 ns this residue underwent several major transitions in backbone dihedral angle. This is consistent with previous studies of glycine dipeptide dynamics, which also found frequent glycine conformational transitions (Hermans et al., 1990).

Side-Chain plus Backbone Clusters for the Extd Trajectory. From the 12 side-chain plus backbone dihedral angle clusters of the extd conformation (Es1–Es12, data not shown), several side-chain interactions were found. To summarize these interactions, in the E1 and E5 conformations, Tyr packs against the peptide backbone, a potentially stabilizing interaction. In the E4 cluster, Pro packs against Val, and in the E2 and E3 conformations, Tyr and Pro occasionally interacted, similar to the extended conformations in the turn trajectory.

Clustering Both Trajectories. It is of interest to determine whether the turn trajectory and the extd trajectory visited similar areas of conformational space. Thus, we compared the conformations by clustering the combined trajectories. T1, T2, and T3 clustered separately, as did E1, E4, and E5. E2 and T5 clustered together, which is not surprising due to their similar conformations (Tables I and II). E3 and T4 also clustered together, with the average conformation between that of the two clusters. Thus, the trajectories intersected in conformational space only for the extended structures.

Comparison of Extd Cluster Results to Those of Tobias et al. We compared the extd clustering results with those of Tobias et al. (see the time line in Figure 7). Tobias et al. observed that the peptide was extended for the first 500 ps. This encompasses both E1/E5 and E3, which had significantly different dihedral angles but were both relatively extended. At 530 ps, Tobias et al. observed a shift from extended to "L-shaped" structures and then "U-shaped" structures, which corresponds to the shift from E3 (extended) to E4 (L-shaped) and then E2 (U-shaped) (Figure 7). Thus, the E2 and T4 clusters, whose conformations clustered together when both trajectories were used, each corresponded to U-shaped structures described by Tobias et al. (see inset, Figure 6).

Tobias et al. described a type II turn between residues 1 and 4 which briefly occurred at 570 ps; this conformation was assigned to the relatively extended E2 cluster. The peptide was fully extended again after 670 ps (agreeing with the shift to E3 observed at this time). This constitutes the first set of transitions observed in the extd trajectory, with the clustering results corresponding well to the full trajectory observations, except for the detection of the type II turn. This will be discussed further below.

A second set of conformations were noted by Tobias et al. in the second part of the trajectory. At approximately 1500 ps, a chain reversal began to occur, with a distorted type I turn involving residues 2–5 occurring at 1530 ps. The turn lasted until 1800 ps, at which time the peptide again adopted an L-shaped conformation. Proline and valine were observed to pack against one another in the type I turn conformation, potentially stabilizing the turn. In the clustering results, the conformations shifted from the E3 to the E4 cluster at approximately 1500 ps (E4 is again the L-shaped cluster noted above). The distorted type I turn, assigned to the E4 cluster, was not resolved as a separate conformation. As in the Tobias et al. study, the Pro/Val packing was observed for the E4 conformations.

Except for the transient type II and type I turns in the extd trajectory, the clustering results described the observed events. Why were the turn conformations not detected? There were several reasons. First, both turns are quite nonideal [see Figures 3 and 5 of Tobias et al. (1991a)], with the last dihedral angle of the four that describe a turn deviating significantly from ideality ($\psi_3 = -120°$ for the type II turn, and $\psi_4 = 143°$ for the type I turn, instead of the ideal 0° found for both turn types). This precludes the formation of a turn hydrogen bond. Turns closer to ideality would have clustered separately due to the radius constraint. Nonetheless, these conformations do represent chain reversals, whereas their associated clusters are fairly extended.

In addition to nonideal dihedral angles, the averaging that takes place during the clustering can also obscure conformations that deviate from the bulk of the conformations. The type II turn, which occurred very briefly, was an "outlier" in the cluster, being over 120° distant from the cluster center (only four conformations were more distant from their cluster's center). This suggests that if transient conformations are of interest, one could cluster conformations to remove the most common, focusing only an outlier conformations.

Unlike the type II turn, the type I turn had dihedral angles quite similar to its cluster center (T4). This points up a concern when dihedral angles are used to describe an extended structure. Small changes in ϕ or ψ near the center of an extended conformation can cause rather large changes in the distance of one peptide end from the other, such as could occur during a chain reversal. Since clustering based on interatomic distances may differentiate such conformations, we used these conformational parameters to cluster each trajectory (data not shown). We found virtually identical clusters for the turn trajectory but significantly different clusters for the extd trajectory. Conformations from E1, E5, and E3 fell into the same clusters, even though their hydrogen-bonding groups were in quite different orientations. The pseudo-type II turn, however, clustered separately, unlike in the dihedral angle clustering. Thus, the differences between conformational parameters must be taken into account when clustering. Both interatomic distance and dihedral angle clustering, however, clustered the type I turn and the L-shaped conformations together. Thus, the "type I turn" described by Tobias et al. may be better classified as a member of the family of L-shaped structures rather than as a separate conformation type.

CONCLUSIONS

The goal of statistical clustering methods is to find a "natural" clustering, though in general this is difficult to define. The weaker requirement is to find clusters that give useful information. In our application, a possible natural clustering would be to group together those conformations that occupy the same basin on the free energy surface. Since the free energy surface could not be calculated from these data (and would require a substantial amount of additional sampling), we apply the weaker condition.

The resultant clusters appeared to separate conformations that differed in an energetically significant way—different hydrogen-bonding patterns or orientations of hydrogen-bonding groups and different van der Waals packing or electrostatic interactions. The clusters also separated conformations that differed by relatively large dihedral angle transitions. Such conformational differences would be expected to fall into different free energy basins.

The resolution of the clusters obtained obviously depended on our choice of a cluster radius. The small number of resulting clusters found here, using a rather large radius, was found to preserve most of the information described by Tobias et al. when analyzing the complete trajectories. The clusters also

gave a useful classification system for studying transitions and hypothesizing what mechanisms caused or allowed these transitions.

When the transitions were examined in detail, we found that most transitions involved a change in a single dihedral angle, though large changes in as many as three angles were also noted. The most common concerted change was the peptide flip, involving anticorrelated changes in ψ_i and ϕ_{i+1}. In both 2-ns trajectories, glycine underwent extensive conformational changes, while only single conformational transitions occurred during the trajectories for the other amino acids.

Several additional points were learned from the observed transitions. Concerted dihedral angle changes were a function of peptide conformation, as changes in angle fluctuations correlated with changes in conformation. Side chains can influence the folding pathway, possibly lowering or raising free energy barriers so different paths through ϕ, ψ dihedral angle space are more or less likely. In the trajectories examined, side-chain and backbone conformational changes were not tightly coupled.

The YPGDV peptide unfolding simulation suggests that the type II turn goes through a solvent-inserted intermediate, similar to unfolding events noted for α helices (Soman et al., 1991; Sundaralingam & Sekharudu, 1989; Tirado-Rives & Jorgensen, 1991; Tobias & Brooks, 1991b). The type II turn adopted a conformation that allowed greater solvent accessibility to the hydrogen bond prior to the unfolding event. This is consistent with studies showing that the free energy barrier to breaking a peptide–peptide hydrogen bond decreases with exposure to water (Sneddon et al., 1989). Interactions which protect the hydrogen bond from solvent exposure (for example, hydrophobic clusters) would be expected to stabilize the type II turn.

Clustering conformations from the dynamics trajectories resulted in meaningful groupings that allowed us to examine the evolution of conformations in a complex trajectory. We can now look for "consensus" interactions within the resulting clusters that point to those peptide–solvent interactions important for facilitating the dihedral transitions. Thus, the sets of similar conformations resulting from clustering trajectory data provide a convenient framework for further analyzing and assimilating the data.

ACKNOWLEDGMENT

We gratefully acknowledge the National Science Foundation for a generous grant of computer time at the Pittsburgh Supercomputing Center. We also thank Dr. J. Mertz of Cray Research Inc. for his assistance in carrying out the molecular dynamics simulations.

REFERENCES

Adzhubei, A. A., Eisenmenger, F., Tumanyan, V. G., Zinke, M., Brodzinski, S., & Esipova, N. G. (1987) *J. Biomol. Struct. Dyn. 5*, 689–704.

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

Brooks, C. L., III, & Karplus, M. (1989) *J. Mol. Biol. 208*, 159–181.

Carpenter, G. A., & Grossberg, S. (1987) *Appl. Opt. 26*, 4919–4930.

Czerminski, R., & Elber, R. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 6963–6967.

DiCapua, F. M., Swaminathan, S., & Beveridge, D. L. (1990) *J. Am. Chem. Soc. 112*, 6768–6771.

Dobson, C. M. (1991) *Curr. Opin. Struct. Biol. 1*, 22–27.

Duda, R. O., & Hart, P. E. (1973) *Pattern Classification and Scene Analysis*, Wiley, New York.

Dyson, H. J., Rance, M., Houghten, R. A., Wright, P. E., & Lerner, R. A. (1988) *J. Mol. Biol. 201*, 161–200.

Everitt, B. (1980) *Cluster Analysis*, Wiley, New York.

Hermans, J., Yun, R.-H., & Anderson, A. G. (1990) in *Crystallographic and Modeling Methods in Molecular Design* (Bugg, C. E., & Ealick, S. E., Eds.) pp 95–113, Springer-Verlag, New York.

IMSL (1987) *STAT/library: FORTRAN Subroutines for Statistical Analysis: User's Manual/IMSL*, IMSL, Houston, TX.

Jorgensen, W. L., Chandrasekhar, J., Madura, J., Impey, R. W., & Klein, M. L. (1983) *J. Chem. Phys. 79*, 926–935.

Karpen, M. E. (1991) Recurring Local Conformations of Protein Structure, Ph.D. Thesis, Case Western Reserve University, Cleveland, OH.

Karpen, M. E. (1992) *Protein Sci.* (in press).

Kim, P. S., & Baldwin, R. L. (1990) *Annu. Rev. Biochem. 59*, 631–660.

Lazaridis, T., Tobias, D. J., Brooks, C. L., III, & Paulaitis, M. E. (1991) *J. Chem. Phys. 95*, 7615–7625.

Levitt, M. (1983) *J. Mol. Biol. 168*, 621–657.

McKelvey, D. R., Brooks, C. L., & Mokotoff, M. (1991) *J. Protein Chem. 10*, 265–271.

Pao, Y.-H. (1989) *Adaptive Pattern Recognition and Neural Networks*, Addison-Wesley, New York.

Polinsky, A., Goodman, M., Williams, K. A., & Deber, C. M. (1992) *Biopolymers 32*, 399–406.

Richardson, J. S., & Richardson, D. C. (1989) in *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G. D., Ed.) pp 1–98, Plenum Press, New York.

Richardson, J. S., Getzoff, E. D., & Richardson, D. C. (1978) *Proc. Natl. Acad. Sci. U.S.A. 75*, 2574–2578.

Rooman, M. J., Rodriguez, J., & Wodak, S. J. (1990) *J. Mol. Biol. 213*, 327–336.

Schellman, C. (1980) in *Protein Folding* (Jaenicke, R., Ed.) pp 53–61, Elsevier/North-Holland Biomedical Press, Amsterdam, The Netherlands.

Sneddon, S. F., Tobias, D. J., & Brooks, C. L., III (1989) *J. Mol. Biol. 209*, 817–820.

Soman, K. V., Karimi, A., & Case, D. A. (1991) *Biopolymers 31*, 1351–1361.

Sundaralingam, M., & Sekharudu, Y. C. (1989) *Science 244*, 1333–1337.

Takane, Y., Young, F. W., & De Leeuw, J. (1977) *Psychometrika 42*, 7–67.

Tirado-Rives, J., & Jorgensen, W. L. (1991) *Biochemistry 30*, 3864–3871.

Tobias, D. J., & Brooks, C. L., III (1991b) *Biochemistry 30*, 6059–6070.

Tobias, D. J., Mertz, J. E., & Brooks, C. L., III (1991a) *Biochemistry 30*, 6054–6058.

Unger, R., Harel, D., Wherland, S., & Sussman, J. L. (1989) *Proteins 5*, 355–373.

Zimmerman, S. S., Pottle, M. S., Nemethy, G., & Scheraga, H. A. (1977) *Macromolecules 10*, 1–9.