JCTC Journal of Chemical Theory and Computation

Automated Event Detection and Activity Monitoring in Long Molecular Dynamics Simulations

Willy Wriggers, Kate A. Stafford, Yibing Shan, Stefano Piana, Paul Maragakis, Kresten Lindorff-Larsen, Patrick J. Miller, Justin Gullingsrud, Charles A. Rendleman, Michael P. Eastwood, Ron O. Dror, and David E. Shaw*

D. E. Shaw Research, New York, New York 10036

Received May 7, 2009

Abstract: Events of scientific interest in molecular dynamics (MD) simulations, including conformational changes, folding transitions, and translocations of ligands and reaction products, often correspond to high-level structural rearrangements that alter contacts between molecules or among different parts of a molecule. Due to advances in computer architecture and software, MD trajectories representing such structure-changing events have become easier to generate, but the length of these trajectories poses a challenge to scientific interpretation and analysis. In this paper, we present automated methods for the detection of potentially important structure-changing events in long MD trajectories. In contrast with traditional tools for the analysis of such trajectories, our methods provide a detailed report of broken and formed contacts that aids in the identification of specific time-dependent side-chain interactions. Our approach employs a coarse-grained representation of amino acid side chains, a contact metric based on higher order generalizations of Delaunay tetrahedralization, techniques for detecting significant shifts in the resulting contact time series, and a new kernel-based measure of contact alteration activity. The analysis methods we describe are incorporated in a newly developed package, called *TimeScapes*, which is freely available and compatible with trajectories generated by a variety of popular MD programs. Tests based on actual microsecond time scale simulations demonstrate that the package can be used to efficiently detect and characterize important conformational changes in realistic protein systems.

1. Introduction

As progress in computer technology has extended the reach of molecular dynamics (MD) simulations[1,2] from picoseconds to nanoseconds and microseconds, complex and functionally important biomolecular motions, such as protein folding and ligand binding, have become more accessible, but the resulting data sets have become increasingly large and unwieldy. Routine MD simulations currently generate trajectories consisting of thousands or millions of frames, rendering both visual inspection and data analysis difficult and time-consuming. We expect that, over time, the analysis of these trajectories will require increasing automation, with human intervention limited to selected events of scientific interest.

We are particularly interested in the detection of significant secondary or tertiary structure rearrangements of proteins, as these motions are often of functional importance. Examples of such large-scale motions include allosteric conformational transitions and folding processes, which give rise to substantial alterations in the interactions between amino acid residues. To shed light on such phenomena, the work described in this paper focuses largely on the automated recognition of significant amino acid contact changes in MD trajectories and on measurement of the *activity*, the total number of such changes per unit time.

Our approach makes use of a particular type of "coarse-grained" model to reduce the level of detail in the spatial

* Corresponding author phone: (212) 478-0260; fax: (212) 845-1286; e-mail: David.Shaw@DEShawResearch.com. David E. Shaw is also with the Center for Computational Biology and Bioinformatics, Columbia University, New York, New York 10032.

10.1021/ct900229u CCC: $40.75 © 2009 American Chemical Society
Published on Web 09/04/2009

representations of long MD trajectories. In particular, we employ a coarse-grained model based on side chains, which offers certain advantages over models based on α-carbon atoms in the context of the present application. Three-dimensional protein structures are described using a distance matrix representation[3,4] that records all pairwise distances between the coarse-grained side chains. In contrast with traditional methods based on the use of global root mean square deviation (rmsd) measurements, the use of distance matrices does not require the translational and rotational alignment of protein structures and facilitates the identification of local structural differences.[5] Our approach decomposes structural changes into a set of key side-chain motions, providing greater sensitivity to a wide range of significant conformational changes than is typically obtained from traditional rmsd-based metrics.

We introduce two alternative approaches to identifying time-dependent contact graphs from distance matrices: a method based on distance cutoffs, which proves useful for detecting local contact formation and breaking activities, and an approach based on Delaunay tetrahedralization, which is better suited to the detection of global folding activities. A recrossing filter is used to eliminate transiently appearing or disappearing edges in the contact graph that are likely to represent random fluctuations and not biologically significant conformational changes.

In the remainder of this paper, we describe the essential elements of our approach, using four microsecond-scale simulations for illustrative purposes. For each trajectory frame, we construct a graph representing all contacts between amino acid side chains, computed using a spatially coarse-grained representation. We track changes in this graph over time, employing a median filter and a recrossing filter for the counting of discrete events that are reflected in the time-dependent contact graph. Finally, we use a kernel measure to derive activity levels from the event data. Although in this work we only examine protein trajectories, it should be relatively straightforward to generalize our approach to, for example, nucleic acids or carbohydrates.

2. Methods

2.1. Molecular Dynamics Simulations. We applied our algorithms to four all-atom MD trajectories, each approximately 1 μs in length. (More detailed system parameters are given in the Supporting Information.) Using the traditional metric of α-carbon rms deviation from the known atomic structure, Figure 1 shows distinct dynamic behavior among the chosen trajectories, which we found particularly useful for method validation. Trajectory 1 (blue) results from a 0.52 μs simulation of Src kinase. In this "generic" trajectory, the system experiences a series of conformational changes, forcing it to increasingly higher rms deviations of up to 4 Å. The "stationary" trajectory 2 (red) corresponds to a stable 1.0 μs simulation of the fast-folding triple mutant K65(NLE), N68H, K70(NLE) of chicken villin subdomain HP-35,[6] where the system remains close to the initial conformation (rms deviation ~1 Å) over the entire length of the simulation. The "diffusive" trajectory 3 (black) shows

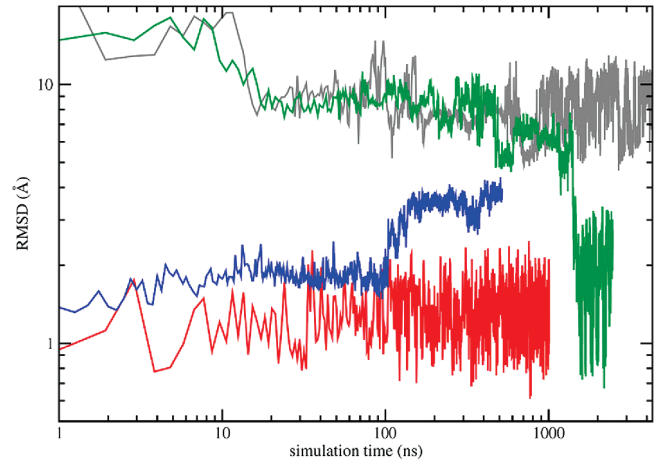

Figure 1. α-Carbon rms deviation from the native conformation as a function of simulation time: trajectory 1 (blue, Src kinase); trajectory 2 (red, villin near-native); trajectory 3 (gray, villin unfolded); trajectory 4 (green, villin folding). The double logarithmic plot bridges between the various time and spatial scales explored by the four trajectories.

the opposite behavior. Starting from an extended (unfolded) villin chain, the system remains far from the native structure during the full 4.3 μs simulation time, visiting various unfolded conformations. Finally, trajectory 4 (green) corresponds to a 2.5 μs "folding" simulation of villin. Together, the four trajectories in Figure 1 cover several scenarios that are commonly encountered in MD simulations of folded and unfolded proteins.

2.2. Coarse-Graining of Side-Chain Contacts. Noncovalent interactions between side chains, such as hydrogen bonds or salt bridges, play a critical role in protein dynamics. Singh and Thornton have shown that each of the 400 possible amino acid side chain pairings exhibits a pronounced peak in its separation histogram at a distance of 5−8 Å.[7] Following this finding, we identify a *representative atom* in each side chain for an efficient calculation of such contact separation distances. (Some side chains have more than one functional group, but our current Delaunay tetrahedralization approach relies on the choice of one representative atom per side chain.) For most residues, we define the second heavy (non-hydrogen) atom counted from the end of the chain as the representative atom. This rule takes into account the fact that in branched residues (e.g., Gln, Asp, or Arg) the end of the chain may be ambiguous, whereas the second heavy atom is straightforward to define in 14 amino acids. Of the remaining six, three aromatic residues (His, Phe, and Tyr) form special cases due to the presence of an aromatic ring; here we pick the atom at the base of the ring (closest to the main chain) as representative. The rare Trp is represented by the epsilon-2 carbon at the center of the double ring. Finally, the achiral Gly and cyclic Pro do not have extended side chains. We represent them by the α- and γ-carbons, respectively, to account for all amino acids.

The idea of reducing the level of detail is not unique to our work, and a number of similar concepts have already been described.[8−13] One possibility is to consider the hydrogen bonding network[14] as a coarse representation of relevant contacts. We have decided against using hydrogen

bonds because they tend to be very transient in MD simulations[15] and provide too much detail; it is often sufficient to know which amino acids are interacting. Another possibility is to select the centroids of side chains[16] or the α-carbon atoms instead of the representative atoms introduced above. Due to the widely variable sizes of side chains, however, the centroids or α-carbons are imprecise markers for interactions with neighboring residues. Alternatively, one could consider the five to seven spatial contact patterns discovered by Singh and Thornton for each of the pairings of amino acids in their *Atlas of Side-Chain Interactions*,[7] but the enumeration of such patterns for every amino acid candidate pair in every trajectory frame would be much more expensive than our simple distance metric.

Given a coarse representation of the structure, an important step in our analysis is to estimate the time-dependent contact pattern (or graph) that captures interactions between representative side-chain atoms. Any such graph approximates the actual atomic interactions, so in practice we can expect some inaccuracies in the assignment of contacts. We introduce two possible approaches for identifying contact graphs with this model: the distance cutoff and the so-called generalized masked Delaunay (GMD) tetrahedralization. Each of these graph-based concepts has its unique advantages for event detection and activity monitoring. While the distance cutoff approach is more selective with respect to local proximity relationships, which is useful for tasks such as distinguishing between the formation and breaking of contacts, the GMD approach accounts for global geometric changes and offers a way to monitor the overall structural variability. For the assignment of the contact graph, it is useful to consider advantages and limitations of these concepts in more detail. Figure 2 provides a schematic overview of proximity measures in two dimensions (the generalization to three dimensions is straightforward). The initial side-chain model is depicted in Figure 2A.

2.3. Distance Cutoff. The cutoff-based metric is the most basic proximity criterion. Contacts are based on the Euclidean distance between representative atoms, and atoms closer than a given cutoff are considered in contact. Parts B and C of Figure 2 illustrate the difficulties associated with identifying contacts by a cutoff distance. If the cutoff is too short (Figure 2B), some valid contacts may be missed, producing false negatives. If the cutoff is too long (Figure 2C), too many undesired contacts are included in the graph, leading to false positives. Such redundant graph edges are typically inconsistent with the actual nearest-neighbor interactions of side chains. In practice, a compromise between these two extreme cases must be found by adjusting the distance cutoff.

The acceptable tolerance for false positives or negatives depends on the application. For example, in α-carbon-based elastic network models, which exhibit a level of detail similar to our side-chain model, the tolerance for false positives is high. Hence, long cutoff distances of 10−15 Å are typically applied in elastic networks, about twice the separation of adjacent α-carbons.[17] Ideally, however, we select in our coarse model only those contacts that correspond to atomic contacts between side chains, requiring us to use a shorter cutoff and leading to a risk of false negatives (Figure 2B) in the resulting contact graph. The distance cutoff criterion also assumes that the side chains are densely packed and that the packing density remains invariant, which is true only for tightly folded proteins.

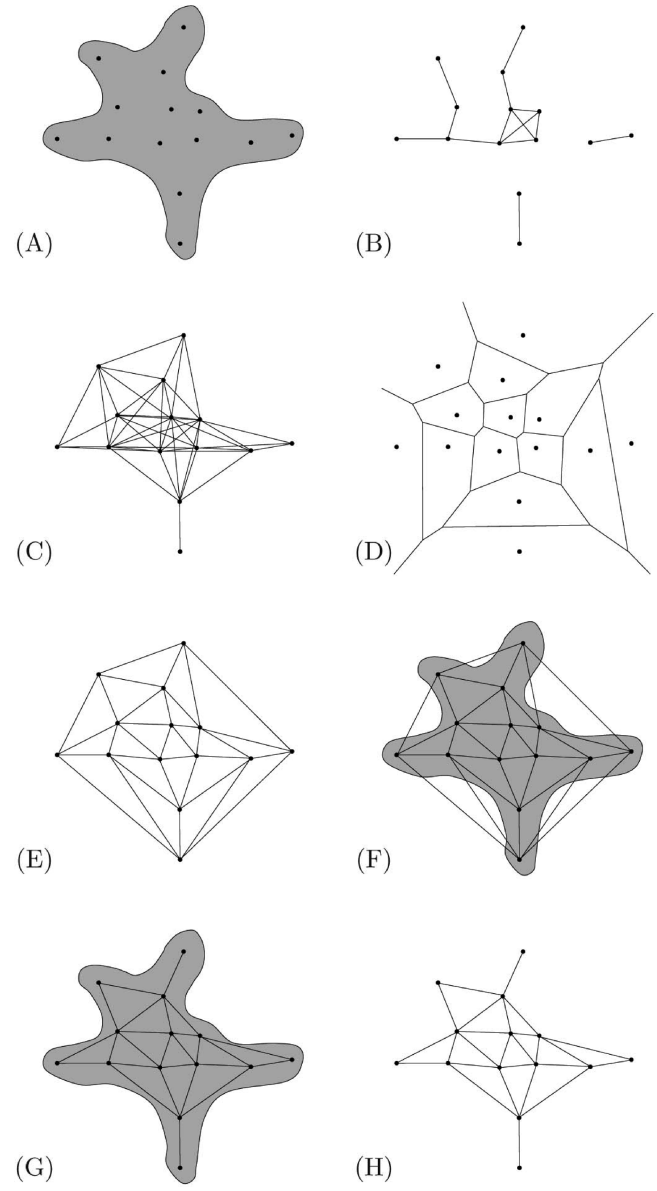

Figure 2. Idealized depiction of computational geometry concepts: (A) Coarse model (black, representative side-chain atoms) superimposed over the "protein" (gray); (B) contacts selected by distance cutoff (too short); (C) contacts selected by distance cutoff (too long); (D) Voronoi cells; (E) Delaunay triangulation; (F) Delaunay triangulation superimposed over protein; (G) masked Delaunay triangulation graph superimposed over protein; (H) masked Delaunay triangulation graph.

2.4. Generalized Masked Delaunay Tetrahedralization. The Voronoi diagram (Figure 2D) and the related Delaunay triangulation (Figure 2E) are well-known proximity measures that automatically adapt to the packing density and do not require cutoff parametrization. Voronoi cells correspond to a nearest-neighbor tessellation of the embedding space:[18] each Voronoi cell contains one representative atom (representing a single side chain) and the region of space that is closer to that representative atom than to any other. A

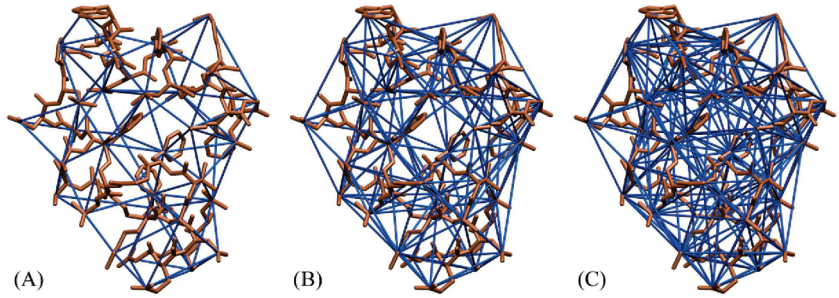

Figure 3. General masked Delaunay (GMD) tetrahedralization of side-chain contacts (blue) in villin (brown, PDB entry 2F4K; see text): (A) Order 2 contacts; (B) order 3 contacts; (C) order 4 contacts. Molecular graphics were created with VMD.[27]

Delaunay graph is the "dual graph" of the Voronoi graph for the same set of representative atoms; one obtains the Delaunay graph by connecting representative atoms whose Voronoi cells share a face or edge. It is straightforward to generalize the first-order Voronoi cells in Figure 2D to higher order; a second-order cell, for instance, corresponds to the regions of space closest to a particular pair of representative atoms. In general, kth-order cells correspond to regions in space that are closest to a particular k-tuple.[19] Such higher order cells might be very small in size (for a depiction see Figure 2 in ref 19).

The Delaunay graph (Figure 2E) appears to be well-suited for our identification of adjacent contacts among representative atoms in three dimensions, but only a subgraph of the Delaunay graph is embedded in the protein structure (schematically shown in Figure 2F). We thus use the so-called "masked Delaunay" tetrahedralization introduced by Martinetz[20] to represent the protein shape more accurately (Figure 2G,H). The Martinetz masking algorithm takes advantage of a theorem (theorem 3 in ref 21) stating that the existence of a second-order Voronoi cell between two representative atoms is equivalent to the presence of a Delaunay edge between them (Figure 2E). An edge-defining second-order Voronoi cell is identified when it contains at least one point of a discretely sampled masking manifold. In our application, the proposed mask is the protein structure and the required discrete sampling is provided naturally by the protein atoms. Figure 3A illustrates the three-dimensional masked Delaunay tetrahedralization for villin.

The original masked Delaunay approach identifies a second-order graph (Figure 3A), connecting pairs (1-simplices) of adjacent representative atoms. We generalize the masked Delaunay approach to higher order, connecting triangles (Figure 3B), tetrahedra (Figure 3C), or, in general, $(k-1)$-simplices, where k is the order of the generalized masked Delaunay graph. This higher order generalization is motivated by the need for a discrete metric for the separation of arbitrary pairs of representative atoms in the GMD context; we use as a metric the minimum order k of the GMD graph for which the pair forms an edge. This discrete k-metric enables us to establish a recrossing filter for accurate detection of contact transitions (further discussed below). The recrossing filter aims to suppress any time-dependent spurious variations in the graph and will also suppress the effect of sampling granularity, i.e., the spacing of generic atoms in the system that might lead to missing GMD edges. To our knowledge, the GMD graph is a new concept, but the related Voronoi cells have already been generalized to higher order, as described above.

Following Martinetz's original definition of the masked Delaunay graph,[20] and sampling the protein mask by the full atom representation, we arrive at a compact formulation of the order-k GMD as applied to biomolecular systems:

(i) Begin with the empty graph G, atom positions $\vec{v}_i \in R^3$ $(i = 1, 2, ..., N)$, and representative side-chain atom positions $\vec{v}_j \in R^3$ $(j = 1, 2, ..., M)$.

(ii) For each atom position $\vec{v}_i$, identify a set of k indices $S_i = \{j_1, j_2, ..., j_k\}$ and its complement S_i^C, $S_i \cup S_i^C = \{1, 2, ..., M\}$ with

$$|\vec{v}_i - \vec{v}_{j_1}| < |\vec{v}_i - \vec{v}_{j_2}| < ... < |\vec{v}_i - \vec{v}_{j_k}| < |\vec{v}_i - \vec{v}_j| \quad (j \in S_i^C)$$

(iii) Add the $(k-1)$-simplex with vertices $(\vec{w}_{j_1}, \vec{w}_{j_2}, ..., \vec{w}_{j_k})$ to G; continue with (ii) until all atoms have been explored.

For a general order k, rule (ii) implies that an edge in the GMD corresponds to a nonempty kth-order Voronoi cell, where in our case the nonempty property refers to the sampling by at least one atom in the system. The rule requires only a partial sorting of the $\vec{w}_j$, which can be efficiently implemented with complexity $O(NkM)$ per trajectory frame. The proposed GMD algorithm is efficient since it does not require an expensive geometric construction of Voronoi polyhedra or Delaunay tetrahedra.

The effect of the GMD order k on the pair distance distribution of the representative side-chain atom model is demonstrated in Figure 4. The tail of the distribution arising from the second-order GMD (a subgraph of the traditional Delaunay tetrahedralization) reaches to distances as high as 10 Å. Figure 4 shows that a 10 Å cutoff would be too permissive and would include many higher order (i.e., redundant) contacts. As a trade-off between false positives and false negatives in cutoff-based graphs, we thus recommend cutoff values of ~7 Å, which would include the peak of the second-order GMD and only a small number of third-order GMD contacts.

2.5. Suppressing High-Frequency Motion. MD time series exhibit a considerable amount of fluctuation on short time scales, introducing noise in the conformational analysis. This noise complicates the reliable identification of significant "level shifts" in the distribution of representative atom pair distances that are relevant over longer time scales (the term "level shifts" is used in time series analysis for low-frequency changes of a nonstationary signal[22]). Such shifts are important both for the cutoff and GMD graphs since they affect

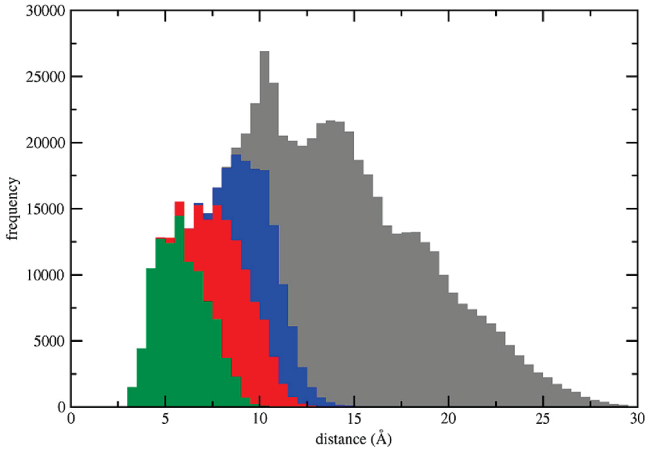

Figure 4. Pair distance distribution histograms for the representative side-chain atom model (see text). Histograms are sorted by the minimum GMD order of a representative atom pair (the smallest number *k* whose GMD includes the edge). Shown in color: minimum GMD order 2 (green), 3 (red), 4 (blue), and >4 (gray). The frequency values were sampled from trajectory 2 (peak distances and shapes of the pair distribution functions are trajectory invariant).

the time-dependent distance matrix and thereby determine the formation and breaking of graph edges. Figure 5A shows a separation distance time series of two representative atoms exhibiting typical level shifts. The two side chains form a contact from 600 to 1400 ns, but a direct assignment of contact formation and breaking using a cutoff of, for example, 8 Å would yield many spurious transitions within this time window due to the noise present on short time scales.

An abundance of alternative low-pass filtering and shift-detection methods have been proposed.[22,23] We tested two well-known and efficient filters for smoothing the time series, the moving average and the median, both defined within a sliding window. In this work, the median is defined as the smallest number in a series such that at least half the numbers are no greater than it. The median is influenced only by the ranking in the sample, making it robust against outliers. The moving average, on the other hand, is a linear filter, and thus it can be easily parallelized if desired. Figure 5B shows the performance of the moving average and median filter as applied to the level shift near 600 ns, indicating that the nonlinear median filter offers a satisfactory preservation of the shift.

In the following section, we implement the median filter for suppressing high-frequency noise in the distance matrix time series. The window half-width, δ, is an important time scale parameter defined by the user which controls the number of events that are detected. In preliminary testing, we have found that half-widths on the order of 10−100 ns provide a reduction of spurious transitions by 2−3 orders of magnitude (Supporting Information Figure 1) relative to the absence of a filter. The choice of δ depends on the time scale of the molecular process investigated by the user.

2.6. Suppressing Trivial Recrossings. One of the well-known problems in transition-state theory[24] is the overcounting of spurious recrossings at the boundary between two states.[12,25] Such recrossings may occur even after median filtering, e.g., in the case of cutoff-based contact graphs when the cutoff is close to the mean of a distance distribution. The "event log" file (see Supporting Information) gives an example of repeated formation and breaking of the same contacts in the absence of any suppression of such recrossings. An overcounting of transitions occurs also for GMD-based events, since the Delaunay tetrahedralization is sensitive to representative atom motions. Several approaches have been proposed to remedy this problem, including the "almost Delaunay" triangulation by Bandyopadhyay and Snoeyink.[26] Here we take a different approach, exploiting the time dependence of the underlying model.

A large number of recrossings is simply an indication that a classification into contacts and noncontacts is not sufficient for the intended purpose of tracking "significant" level shifts. To compensate for these unwanted effects, we have developed a "trivial recrossing suppression" scheme (see Supporting Information). The idea, discussed in the "stable states picture" of chemical reactions[25] and recently used in the construction of Markov models from MD simulations,[12] is to introduce a buffer region and to track crossings until this buffer has been crossed completely. Figure 6 provides an overview of the nine possible paths crossing the buffer and identifies the remaining "nontrivial" contact formation and breaking events (green and red arrows, respectively), after application of the recrossing filter. Numeric labels assigned to the regions by our algorithm (Supporting Information) are also shown.

The use of a buffer requires the definition of a "contact metric" that separates the buffer from contacts and noncontacts. The metric may be continuous, as in the case of cutoffs, or discrete, as in the case of GMD graphs, where we use the minimum GMD order of an edge as metric (the smallest number *k* whose GMD includes the edge). The width of the buffer region is a free parameter defined by the user. In tests using cutoff-based contacts and the stationary trajectory 2 (which exhibits little activity and is thus a good test system for detecting spurious recrossings), we have found that even very small buffer zones of 0.3−0.5 Å are highly effective in eliminating unwanted recrossings (Supporting Information Figure 2). In the case of GMD, we found the smallest possible buffer with a minimum order 2 (contacts), 3 (buffer), and 4 or higher (noncontacts) to be effective; it will be denoted as the "$k = 3$" crossing buffer in the following discussion.

2.7. Kernel-Based Activity Measure. The analysis described so far yields a detailed listing of K broken and formed contacts at corresponding times t_i ($i = 1, ..., K$). The cutoff- or GMD-based activities (rates of events) are computed from the event times by smoothing with a Gaussian kernel:

$$a(t) = \frac{1}{\sqrt{2\pi}\sigma} \sum_{i=1}^{K} e^{-(t - t_i)^2/2\sigma^2} \qquad (1)$$

The activity $a(t)$ is not normalized to unity as in probability density estimation, but to K, the total number of events, such that a gives the number of events per frame. The kernel standard deviation σ is matched to the median half-width δ as follows. The median filter can be considered a low-pass

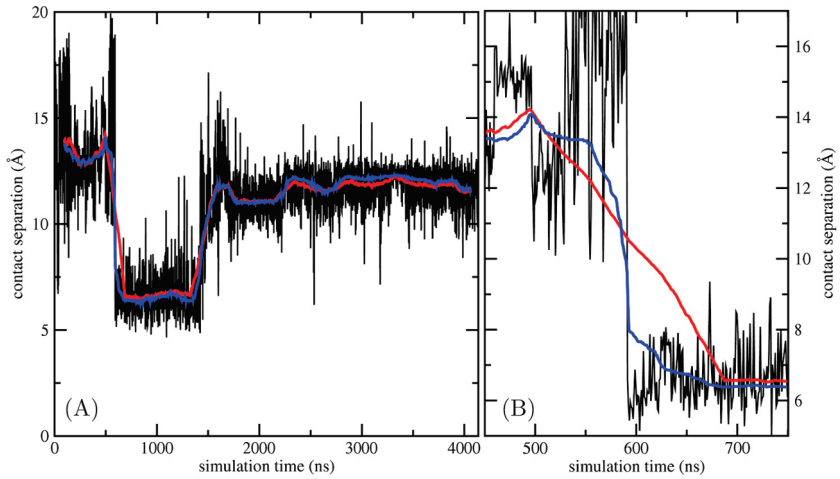

Figure 5. Smoothing of a typical contact time series (black) by moving average (red) and median filters (blue): (A) Full time window; (B) detailed view of a level shift at 600 ns. Shown is the separation of model atoms representing Asp5 and Phe10 in trajectory 3. The moving average and median filters used a sliding window of half-width = 100 ns.

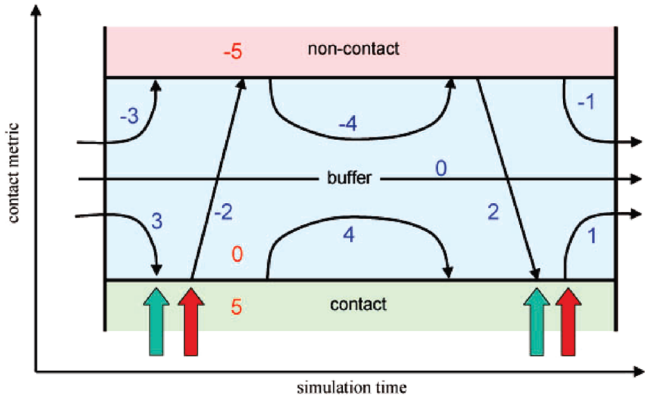

Figure 6. Suppression of trivial recrossings using a buffer zone (blue) between contact (green) and noncontact (red) zones. Nine types of buffer boundary crossings (thin black arrows) are theoretically possible. The colored arrows mark the time of four designated crossings (green, contact formation; red, contact breaking). All other crossings are suppressed (see text). Red numbers show initial numeric values used in the bidirectional tracking (see Supporting Information). Blue numbers are the final labels assigned to each of the nine crossing types. The metric for assigning zone boundaries may be continuous (distance cutoff) or discrete (GMD minimum order).

filter that attenuates frequencies above $(2\delta)^{-1}$ (the inverse of the median window width). The minimum sampling rate (or Nyquist rate) should be twice this frequency, or δ^{-1}, according to the Nyquist−Shannon sampling theorem. The "full width at half-maximum" (fwhm) parameter is commonly used to describe the resolving width of a kernel. This width must be small enough to resolve Nyquist rate samples. For the Gaussian kernel, the fwhm $= 2(2 \ln 2)^{1/2}\sigma$ is thus matched to the inverse Nyquist rate: fwhm $= \delta$.

Although it would, in principle, be possible to sample above the Nyquist rate (i.e., δ could be considered an upper bound for the kernel fwhm), we note that the smoothness of the activity curves is critical for estimating basin minima and basin transitions corresponding to local extrema of $a(t)$, so the maximum fwhm $= \delta$ is chosen in our application to ensure the maximum smoothness of $a(t)$ (see Results and Discussion). The smoothing parameter δ thus corresponds to both the half-width of a median filter and the fwhm of a Gaussian kernel in our application.

2.8. Output. Our implementation provides a number of output files for inspection, plotting, and visualization of the methods described above: (a) a detailed log file of formation and breaking of contacts (for an example, see Supporting Information); (b) an activity time series data file containing the frame number, combined activity $a(t)$, and separate activities derived from either formation or breaking events; (c) trajectory files containing basin minima and basin transitions corresponding to local extrema of the combined activity $a(t)$; (d) a VMD-readable[27] contact graph for each frame (Figure 3A), enabling animation of contact graphs.

In the following section, we illustrate the use of the proposed analysis tools in practical MD applications.

3. Results and Discussion

The major idea associated with the tools introduced in the previous section is their ability to decompose the overall dynamics (expressed by the activity curves $a(t)$ of eq 1) into constituent individual events related to the breaking and formation of amino acid side-chain contacts. Before assessing the utility of detailed event logs in the practical analysis workflow, it is useful to compare the activities $a(t)$ to more traditional rms alignment techniques. Any similarities with the traditional techniques are nontrivial due to the different methodological paths taken by our methods. Differences, on the other hand, will suggest application areas for which our strategies are uniquely specified. We will describe two especially advantageous applications, the visualization of activity measures and the identification of activity basins and transitions in the trajectory.

3.1. Comparison of Tools for Activity Analysis. Figures 7 and 8 show the results of GMD-based (A) and cutoff-based (B) activity analysis applied to the "generic" trajectory 1 and the "folding" trajectory 4 (results for trajectories 2 and 3 are shown in Supporting Information Figure 2 and in Figure 9, respectively). For comparison with traditional

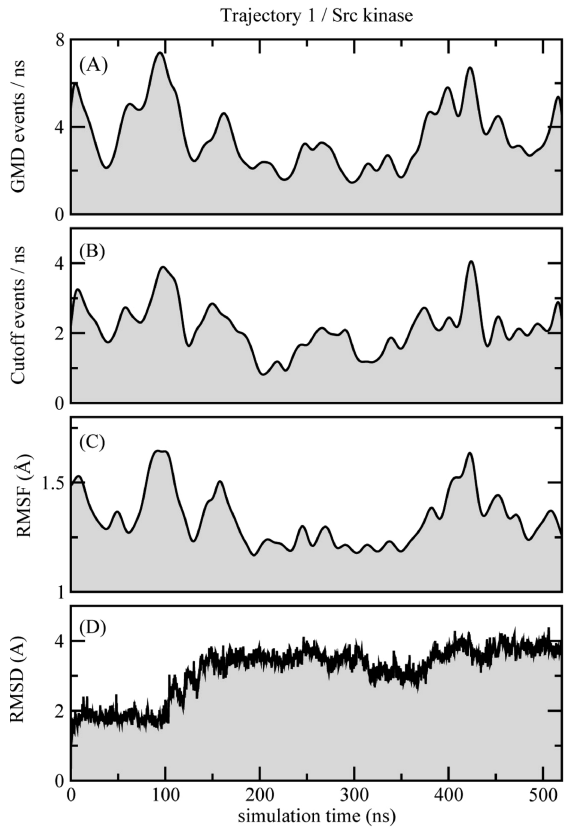

Figure 7. Comparison of conformational analysis tools applied to trajectory 1: (A) GMD-based activity ($k = 3$ crossing buffer); (B) cutoff-based activity (6.0−7.0 Å crossing buffer); (C) rms fluctuation in a sliding window; (D) α-carbon rms deviation from PDB entry 1Y57. The smoothing parameter δ setting Gaussian fwhm and median half-widths (see text) was 12.5 ns.

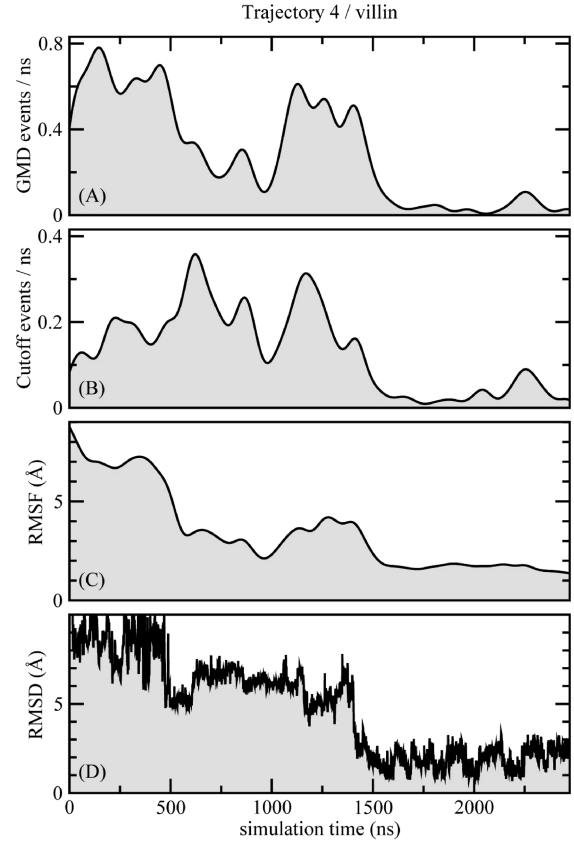

Figure 8. Comparison of conformational analysis tools applied to trajectory 4: (A) GMD-based activity ($k = 3$ crossing buffer); (B) cutoff-based activity (7.5−8.5 Å crossing buffer); (C) rms fluctuation in a sliding window; (D) α-carbon rms deviation from PDB entry 2F4K (with "A" variants of dual occupancy rotamers). The smoothing parameter δ setting Gaussian fwhm and median half-widths (see text) was 100 ns.

techniques, Figures 7 and 8 present also the rms fluctuation (C) and rms deviation from the native structure (D). The rms fluctuation in C measures the all-atom variability of consecutive frames in the trajectory weighted by a sliding Gaussian envelope function. To provide comparable detail, we have again matched the fwhm of the Gaussian envelope to the smoothing parameter δ (see above). We note that the unusual choice of a Gaussian envelope function for smoothing the rms fluctuations is critical for allowing comparison between these curves. If we used a more traditional sliding box envelope for the rms fluctuations, the curves in C would exhibit high-frequency noise (not shown), reducing the similarity with those in A and B.

The "generic" Src kinase trajectory 1 in Figure 7 represents a frequently encountered MD scenario and is thus of particular utility for the comparison of analysis tools. We describe similarities of analysis techniques by the Pearson correlation coefficient. The GMD-based (A) and cutoff-based activities (B) are quite similar in this case (correlation 0.86). Likewise, both activities are similar to the time-dependent rms fluctuation (C; correlations 0.90 and 0.84 for GMD- and cutoff-based activity, respectively). It is reassuring that the three measures (Figure 7A−C) are consistent in their characterization of traditional MD trajectories, even though there are considerable methodological differences in their design. A minor difference from the two activity measures is the elevated background level exhibited by the rms fluctuation (Figure 7C), but this is inconsequential for analysis. Differences are more pronounced when comparing the three measures (Figure 7A−C) to the rms deviation. The first three measures show increased activity preceding a pronounced conformational change evident in the rms deviation (Figure 7D) after 100 ns (see also below). The subsequent activity peaks are not seen to have any major effect on the rms deviation. For example, the peak at 420 ns is due to local fluctuations in the disordered C-terminus which do not affect the rms deviation, since the structure has already moved far from the native conformation at this point.

The villin folding trajectory 4 in Figure 8 offers an opportunity to analyze a trajectory going from an extended to a compact, native state. The rms deviation (D) shows that the protein folds at 1400 ns. The GMD-based (A) and cutoff-based (B) activities yield a more detailed picture of the dynamic activity of the system up to 1400 ns, although the measures exhibit striking differences in this case (correlation 0.64). The major difference at the beginning of the trajectory is due to the fact that most contacts are outside the cutoff range in the initial extended conformation, but such folding events are included in the GMD, which does not depend on

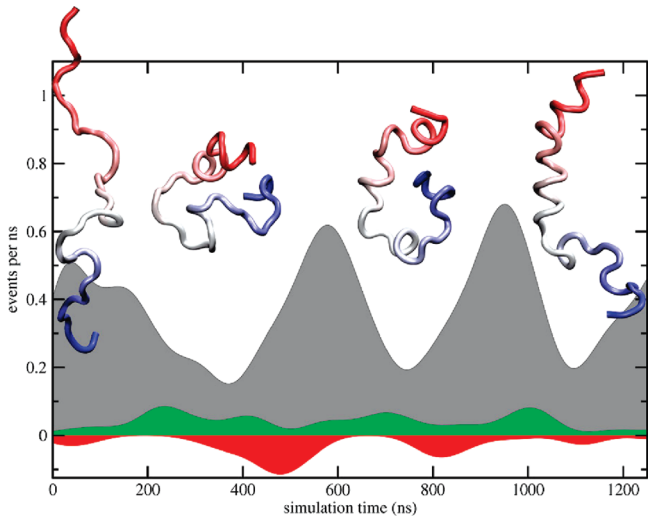

Figure 9. Activity levels exhibited by the diffusive trajectory 3 during the first 1.25 μs: total GMD activity (gray); cutoff-based contact formation activity (green); cutoff-based contact breaking activity (red, plotted in negative direction to simplify comparison). A median filter (see text) was applied, using a half-width of 100 ns. Recrossing suppression used a buffer of 6.0−7.0 Å (cutoff) or in the case of GMD, a buffer of $k = 3$ (see text). Snapshots of the trajectory above the plot correspond to the initial conformation and to three local minima of GMD activity that represent the basins directly below them. Molecular graphics renderings were created with VMD.[27] An animated AVI version of this figure, showing the full length of the trajectory, is available in the Supporting Information.

the cutoff. This difference highlights the adaptive property of contacts employed by the GMD approach. The rms fluctuation (C) is more similar (correlation 0.91) to the GMD- than to the cutoff-based activity (correlation 0.46). Despite the relatively high correlation, the local variations of the rms fluctuation (C) are significantly attenuated in this example compared to the variations of the GMD activity (A), in contrast to Figure 7, where the two measures show similar variability. Also, a small activity peak at 2200 ns is missed by the rms fluctuation. As described in subsection 3.2, this activity peak corresponds to substantial fluctuations of the first helix.

Our results suggest that the rms deviation is the least reliable predictor of conformational transitions because it misses some events detected by the other measures once the rms deviation reaches high numeric values. Also, GMD- and cutoff-based activities provide some additional information one could not obtain from rms fluctuations. The GMD- and cutoff-based activities differ especially in the folding trajectory 4. We explore differences between GMD- and cutoff-based activities further in subsection 3.3.

3.2. Utility of Detailed Event Logs. One important advantage of the proposed analysis is that it provides a detailed listing of constituent events that facilitates an underlying structural interpretation of the activity, beyond detection of periods of high activity itself. Traditional analysis tools based on Cartesian coordinates are not able to provide such detail. The cases of trajectories 1 and 4 illustrate the utility of event logs provided by the new algorithms. These can be particularly useful when combined with expert knowledge, for example from mutagenesis data, of which residues are believed to play an important role.

The event logs of the Src kinase trajectory 1 indicate that Phe405 undergoes a conformational change that results in its exchange of packing partner during the 90−140 ns time period. Initially, Phe405 is in proximity of Glu310, Val313, Leu317, and Met314, contacts which are broken at 89, 90, 96, and 136 ns, respectively. The loss of contacts is compensated by the formation of a new contact with His384 at 96 ns simulation time. This conformational change mainly involving Phe405, His384, and Met314 is highly intriguing and potentially important, since Phe405 and His384 belong to the well-known DFG and HRD motifs that are almost universally conserved among protein kinases, and Met314, Phe405, and His384 are all part of a critical structural "spine" that was identified to stabilize kinase active structures.[28]

For the villin folding trajectory 4, formation of helical (i,i+3) and (i,i+4) contacts contribute substantially to the activity in the initial part of the trajectory. This is not directly followed by folding, but rather the protein appears temporarily trapped due to the formation of nonnative interactions. Specifically, after approximately 600 ns, a contact forms between the oppositely charged N- and C-terminal residues. This contact, together with an overextension of helix 1 through to residue Thr13, characterizes a persistent nonnative state between approximately 900 and 1100 ns that is associated with a dip in activity (Figure 8). Exit from this state is accompanied by the loss of the nonnative helical contacts in helix 1 and the subsequent formation of helix 2. The final event in folding is the unraveling and re-formation of helix 1, together with a reorientation of the loop between helices 1 and 2. This is accompanied by a burst of contact formation between hydrophobic residues, including the Phe6-Phe17 contact in the core, which is formed at ∼1400 ns. After folding, helix 1 occasionally undergoes substantial fluctuations, leading to the rise in activity at approximately 2200 ns visible in Figure 8. This involves the partial transient loss of helical structure from helix 1, reflected in changes in the contacts in that helix, accompanied by a change in orientation of helix 1 with respect to the rest of the protein that is reflected in changing contacts between residues at the beginning of helix 1 with those near the beginning of helix 2.

Once the contact formation or breaking events are identified, geometric inspection tools such as those provided by VMD[27] may add to the interpretation. It would have been impossible to extract this highly specific information with one of the traditional rms deviation or rms fluctuation measures.

3.3. Visualization of Activity Results. Given the differences between GMD and cutoff when applied to folding trajectory 4 (Figure 8B), we have investigated the discrepancy further using the "diffusive" trajectory 3. Since the original level shifts that give rise to activities can be separated into formation and breaking events, we considered separately the formation and breaking activities derived from the two classes. The differences were striking for cutoff-based activity levels (Supporting Information Figure 3; correlation 0.07 between formation and breaking), whereas in the case of

GMD the formation and breaking contributions were very similar (Supporting Information Figure 4; correlation 0.77). This indicates that, at least in the case of folding trajectories, the formation and breaking of cutoff-based contacts are asymmetric and at times either one may be dominating, whereas in the GMD graph the total number of contacts is nearly constant. We thus propose to visualize separate formation and breaking activities in the case of cutoff contacts, and only the total activity of the GMD.

Figure 9 displays such a "combination plot" of activities together with snapshots of the trajectory at low GMD activity. An animated AVI version of this visualization is available in the Supporting Information. One can observe at several times the pronounced asymmetry in the cutoff activity levels. A dominant formation of cutoff contacts, such as at 250, 700, and 1000 ns, typically precipitates a stabilization of the system (as judged by low GMD activity at 350, 750, and 1100 ns). Likewise, a dominant breaking of cutoff contacts, such as at 500 and 800 ns, clearly favors subsequent folding transitions (corresponding to high GMD activity at 600 and 900 ns). The proposed combination plot thus provides a nuanced characterization of folding activity, in which periods of stabilization or destabilization of the overall fold can be matched with more detailed changes in the side-chain packing.

The results suggest that inactive periods observed in folding trajectories are caused by preceding periods of contact formation of the structure, whereas large-scale folding transitions follow after periods of contact destabilization. The observed dependence of structural stability on contact formation could be used to enhance sampling in folding trajectories.

3.4. Segmentation of Activity Basins and Transitions. Figure 9 suggests a natural segmentation of the trajectory into quiescent "basins" separated by "transitions." We assigned local minima (basin centers) and local maxima (transitions) using a finite difference approximation of the first derivative of the total (GMD- or cutoff-based) activity $a(t)$. The local maxima correspond to highly active periods of the trajectory that separate basins of inactivity. The local minima roughly correspond to the structures with the greatest contact similarity to the average structure of the local basin. These minima are shown in Figure 9 above the GMD activity plot, representing the inactive basins directly below them. This strategy can also be applied (after Gaussian smoothing) to the traditional rms fluctuation.

For typical MD trajectories such as trajectory 1, the maxima and minima of $a(t)$ are not very sensitive to the graph method used. For example, 75% of the minima and transitions derived from the GMD activity (Figure 7A) can be found to be within 5 ns simulation time of like extrema exhibited by the rms fluctuation (Figure 7C). The similarity with the rms fluctuation was somewhat less pronounced for the cutoff-based activity (63%; Figure 7B). As can be expected, the observed conservation of minima and transitions agrees qualitatively with the above Pearson correlation analysis. We propose to use the GMD activities for the assignment of basins and transitions whenever possible, due to the more pronounced undulations relative to the rms fluctuation (Figure 8A,C).

4. Conclusion

We have introduced tools for automated event detection and activity monitoring in MD simulations and demonstrated their application to state-of-the-art trajectories. Our method introduces intuitive parameters to be defined by the user, as follows:

(a) The type of contact graph. We recommend a cutoff-based graph to detect detailed side-chain contact formation and breaking, or a GMD-based graph to detect global activity.

(b) The designated crossing buffer. We recommend 6−7 Å cutoffs or GMD order $k = 3$.

(c) The temporal smoothing parameter δ. This value depends on the length of the simulation and the desired level of detail.

(d) The side-chain atom selection. We provide a default profile for standard amino acid residues, which may be modified for specific systems or nonstandard residues.

All other steps in the methodology are automated, including median filtering, suppression of trivial recrossings, kernel activity estimation, calculation of basin minima and transitions, and data file output.

Our current serial implementation is sufficiently efficient to allow for the analysis of microsecond-scale trajectories. An analysis of 4496 frames of trajectory 3 took only a few minutes of compute time on a standard Linux workstation. For much longer trajectories, we expect that parallelization of the analysis may be required; such parallelization should be straightforward using, for example, the recently developed HiMach framework.[29]

Our implementation brings together state-of-the-art methodologies from time-series analysis, computational geometry, graph theory, and biochemistry to address the activity-monitoring and event-detection problem. The limitations of our methods include the focus on global rearrangements in the structure; some events of scientific interest leave only very small footprints in the surrounding protein matrix. Ion and solvent diffusion through membrane channels, for example, would require different detection techniques. In addition, the parameters of our method have not yet been optimized for lipids and nucleic acids, although it would in principle be possible to generalize the coarse-grained model to nonprotein contacts—especially in the case of GMD, which is independent of specific cutoff distances.

An additional limitation of our analysis is that events are still relatively frequent for human interpretation (about 100−1000 events were observed per microsecond). For longer trajectories, it may be helpful to further reduce the complexity of the contact patterns using one or more of the following strategies: (i) ignoring contacts formed by residues with nonexistent or short side chains such as Cys, Pro, and Ala; (ii) substantially increasing the crossing buffer; (iii) ranking events by the sequence conservation of participating residues, the energy levels of participating residues, or a correlation analysis of the motion of participating residues.

Our tests on four trajectories have revealed a number of advantages of our activity-based calculations relative to the

more traditional rms fluctuation. These include (i) a higher sensitivity at low activity levels (Figure 8A,C); (ii) a reduced background noise contribution (Figure 7A, C); (iii) a detailed listing of individual events underlying the observed activity; (iv) coarse model calculations that are roughly an order of magnitude faster than an all-atom analysis; and (v) a functionally relevant diversification of the tool arsenal: the GMD activities show an overall fold rearrangement, the cutoff activities measure contact formation and breaking, and the traditional rms fluctuation measures the variability of Cartesian coordinates of neighboring frames. The importance of automated analysis techniques will only grow as efforts in high-throughput MD simulation—such as the "Dynameomics" project[30]—make large numbers of MD trajectories publicly available for mining and interrogation.

4.1. Dissemination. All tools described in this article will be documented and freely distributed as part of the Python-based "TimeScapes" package at URL http://www.DEShawResearch.com (Resources). TimeScapes is capable of reading the trajectories produced by many popular MD programs, including AMBER, CHARMM, NAMD, X-PLOR, Desmond, LAMMPS, and GROMACS, making the package widely applicable.

Supporting Information Available: Supporting methods, supporting figures, an events log file, and an animation in AVI format. This information is available free of charge via the Internet at http://pubs.acs.org.

Acknowledgment. We thank Morten Jensen, Tiankai Tu, and Michael Gross for helpful discussions.

References

(1) Karplus, M.; McCammon, J. A. Molecular Dynamics Simulations of Biomolecules. *Nat. Struct. Biol.* **2002**, *9*, 646–652.

(2) Adcock, S. A.; McCammon, J. A. Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chem. Rev.* **2006**, *106*, 1589–1615.

(3) Havel, T. F.; Kuntz, I. D.; Crippen, G. M. The Theory and Practice of Distance Geometry. *Bull. Math. Biol.* **1983**, *45*, 665–720.

(4) Holm, L.; Sander, C. Structure Comparison by Alignment of Distance Matrices. *J. Mol. Biol.* **1993**, *233*, 123–138.

(5) Keller, P. A.; Leach, S. P.; Luu, T. T.; Titmuss, S. J.; Griffith, R. Development of Computational and Graphical Tools for Analysis of Movement and Flexibility in Large Molecules. *J. Mol. Graphics Modell.* **2000**, *18*, 235–241.

(6) Kubelka, J.; Chiu, T. K.; Davies, D. R.; Eaton, W. A.; Hofrichter, J. Sub-microsecond Protein Folding. *J. Mol. Biol.* **2006**, *359*, 546–553.

(7) Singh, J.; Thornton, J. M. *Atlas of Protein Side-Chain Interactions*, Vols. I and II; IRL Press: Oxford, U.K., 1992.

(8) Yang, H.; Parthasarathy, S. Mining Spatial and Spatio-Temporal Patterns in Scientific Data. *Proceedings of the 22nd International Conference on Data Engineering Workshops*, Atlanta, GA; IEEE Computer Society: Washington, D.C., 2006.

(9) Zhou, R.; Parida, L.; Kapila, K.; Mudur, S. PROTERAN Animated Terrain Evolution for Visual Analysis of Patterns in Protein Folding Trajectory. *Bioinformatics* **2007**, *23*, 99–106.

(10) Schütte, C.; Fischer, A.; Huisinga, W.; Deuflhard, P. A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo. *J. Comput. Phys.* **1999**, *151*, 146–168.

(11) Deuflhard, P.; Huisinga, W.; Fischer, A.; Schütte, C. Identification of Almost Invariant Aggregates in Reversible Nearly Uncoupled Markov Chains. *Lin. Alg. Appl.* **2000**, *315*, 39–59.

(12) Buchete, N.-V.; Hummer, G. Peptide Folding Kinetics from Replica Exchange Molecular Dynamics. *Phys. Rev. E* **2008**, *77*, 030902(R).

(13) Yaliraki, S. N.; Barahona, M. Chemistry Across Scales: From Molecules to Cells. *Philos. Trans. R. Soc. A* **2007**, *365*, 2921–2934.

(14) Factor, A. D.; Mehler, E. L. Graphical Representation of Hydrogen Bonding Patterns in Proteins. *Protein Eng.* **1991**, *4*, 421–425.

(15) Sessions, R. B.; Gibbs, N.; Dempsey, C. E. Hydrogen Bonding in Helical Polypeptides from Molecular Dynamics Simulations and Amide Hydrogen Exchange Analysis: Alamethicin and Melittin in Methanol. *Biophys. J.* **1998**, *74*, 138–152.

(16) Kaźumierkiewicz, R.; Liwo, A.; Scheraga, H. A. Addition of Side Chains to a Known Backbone with Defined Side-Chain Centroids. *Biophys. Chem.* **2003**, *100*, 261–280.

(17) Jeong, J. I.; Jang, Y.; Kim, M. K. A Connection Rule for Alpha-Carbon Coarse-Grained Elastic Network Models Using Chemical Bond Information. *J. Mol. Graphics Modell.* **2006**, *24*, 296–306.

(18) Martinetz, T.; Schulten, K. Topology Representing Networks. *Neural Networks* **1994**, *7*, 507–522.

(19) Fischer, I.; Gotsman, C. Fast Approximation of High Order Voronoi Diagrams and Distance Transforms on the GPU. *J. Graphics Tools* **2006**, *11*, 39–60.

(20) Martinetz, T. Competitive Hebbian Learning Rule Forms Perfectly Topology Preserving Maps. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN-93)*; Gielen, S., Kappen, B., Eds.; Springer Verlag: Heidelberg, Germany, 1993; pp 427−434.

(21) de Berg, M.; van Kreveld, M.; Overmars, M.; Schwarzkopf, O. *Computational Geometry: Algorithms and Applications*; Springer Verlag: Berlin, 2000.

(22) Gather, U.; Fried, R.; Lanius, V. Robust Detail-Preserving Signal Extraction. In *Handbook of Time Series Analysis*; Schelter, B., Winterhalder, M., Timmer, J., Eds.; Wiley-VCH: Weinheim, Germany, 2006; pp 131−157.

(23) Ye, L.; Wu, Z.; Eleftheriou, M.; Zhou, R. Single-Mutation-Induced Stability Loss in Protein Lysozyme. *Biochem. Soc. Trans.* **2007**, *35*, 1551–1557.

(24) Pollak, E.; Talkner, P. Reaction Rate Theory: What It Was, Where Is It Today, and Where Is It Going. *Chaos* **2005**, *15*, 026116.

(25) Northrup, S. H.; Hynes, J. T. The Stable States Picture of Chemical Reactions. I. Formulation for Rate Constants and Initial Condition Effects. *J. Chem. Phys.* **1980**, *73*, 2700–2714.

(26) Bandyopadhyay, D.; Snoeyink, J. Almost-Delaunay Simplices: Nearest Neighbor Relations for Imprecise Points. *Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete*

Algorithms, New Orleans, LA; Society for Industrial and Applied Mathematics: Philadelphia, 2004; pp 410−419.

(27) Humphrey, W. F.; Dalke, A.; Schulten, K. VMD-Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33–38.

(28) Kornev, A. P.; Haste, N. M.; Taylor, S. S.; Ten Eyck, L. F. Surface Comparison of Active and Inactive Protein Kinases Identifies a Conserved Activation Mechanism. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 17783–17788.

(29) Tu, T.; Rendleman, C. A.; Borhani, D. W.; Dror, R. O.; Gullingsrud, J.; Jensen, M. O.; Klepeis, J. L.; Maragakis, P.; Miller, P.; Stafford, K. A.; Shaw, D. E. A Scalable Parallel Framework for Analyzing Terascale Molecular Dynamics Trajectories. *Proceedings of the 2008 ACM/IEEE Conference on Supercomputing*, Austin, TX; ACM Press: New York, NY, 2008.

(30) Beck, D. A. C.; Jonsson, A. L.; Schaeffer, R. D.; Scott, K. A.; Day, R.; Toofanny, R. D.; Alonso, D. O. V.; Daggett, V. Dynameomics: Mass Annotation of Protein Dynamics and Unfolding in Water by High-Throughput Atomistic Molecular Dynamics Simulations. *Protein Eng. Des. Sel.* **2008**, *21*, 353–368.

CT900229U