Published in final edited form as:
Proteins. 2012 November ; 80(11): 2536–2551. doi:10.1002/prot.24135.

Event Detection and Sub-state Discovery from Bio-molecular Simulations Using Higher-Order Statistics: Application To Enzyme Adenylate Kinase

Arvind Ramanathan[1], **Andrej J. Savol**[2,3], **Pratul K. Agarwal**[1], and **Chakra S. Chennubhotla**[3,*]

[1]Computational Biology Institute & Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA, 37830.

[2]Joint Carnegie Mellon University–University of Pittsburgh Ph.D. Program in Computational Biology, Pittsburgh, PA, USA.

[3]Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA, USA, 15260.

Abstract

Biomolecular simulations at milli-second and longer timescales can provide vital insights into functional mechanisms. Since post-simulation analyses of such large trajectory data-sets can be a limiting factor in obtaining biological insights, there is an emerging need to identify key dynamical events and relating these events to the biological function online, that is, as simulations are progressing. Recently, we have introduced a novel computational technique, quasi-anharmonic analysis (QAA) (PLoS One 6(1): e15827), for partitioning the conformational landscape into a hierarchy of functionally relevant sub-states. The unique capabilities of QAA are enabled by exploiting anharmonicity in the form of fourth-order statistics for characterizing atomic fluctuations. In this paper, we extend QAA for analyzing long time-scale simulations online. In particular, we present HOST4MD - a higher-order statistical toolbox for molecular dynamics simulations, which (1) identifies key dynamical events as simulations are in progress, (2) explores potential sub-states and (3) identifies conformational transitions that enable the protein to access those sub-states. We demonstrate HOST4MD on micro-second time-scale simulations of the enzyme adenylate kinase in its *apo* state. HOST4MD identifies several conformational events in these simulations, revealing how the intrinsic coupling between the three sub-domains (LID, CORE and NMP) changes during the simulations. Further, it also identifies an inherent asymmetry in the opening/closing of the two binding sites. We anticipate HOST4MD will provide a powerful and extensible framework for detecting biophysically relevant conformational coordinates from long time-scale simulations.

Keywords

molecular dynamics; anharmonic motions; adenylate kinase; quasi-anharmonic analysis; principal component analysis

1. Introduction

Molecular dynamics (MD) and Monte-Carlo (MC) techniques are widely used tools to investigate conformational changes within proteins [1]. The availability of specialized

*chakracs@pitt.edu.

hardware and optimized codes for graphics processing units (GPUs) has increased the time-scales available to MD simulations (milliseconds and beyond) [2–5]. Longer time scale simulations produce correspondingly larger datasets; repositories such as Dynameomics [6] and Folding@Home [7] contain several petabytes of data (representing more than tens of millions of conformational snapshots) which are content rich but pose serious challenges for analysis and interpretation.

Automating the analysis of long time-scale simulations is an important and recognized challenge faced by the biophysics community [8]. In particular, extracting information from trajectories regarding conformational states that involve substantial structural re-arrangements (and/or dynamical correlations between residues or regions in the protein) is often challenging and a limiting factor in elucidating the biophysical mechanisms of protein function. A number of tools have been proposed to address some of these challenges. Techniques based on principal component analysis (PCA) [9, 10] have been used to extract simplified descriptions of the conformational landscape in the context of protein folding [11], enzyme catalysis [12] and molecular recognition [13]. Non-linear dimensionality reduction techniques have also been used to extract reaction coordinates for describing protein-folding processes [14]. Clustering techniques [15–18], including Markov state models [19] are also commonly employed techniques for organizing large conformational spaces, identifying conformational sub-states, and deriving rate kinetics.

Although post-processing of trajectories is important to gain biological insights, the increase in the size of trajectories from long time-scale simulations necessitates for some of this analyses to be done online, i.e., as the simulations progress. TimeScape [20] is one such recently introduced tool to monitor the time-evolution of inter-residue contacts and performs time-dependent segmentation of the root-mean squared (RMS) fluctuations into ‘basins’ and ‘transitions’ for temporal coarse-graining of long time-scale simulations. Langmead and coworkers [21] introduced dynamic tensor analysis (DTA) to identify flexible/rigid regions in a protein and to extract conformational sub-states using second order statistics (covariance) of atomic fluctuations in real-time. Protein anharmonicity, a well-documented feature of the conformational landscape, has not been addressed in any online method [22–32].

The thermal motions of atoms produce position distributions that have significant high-order moments, suggesting that non-Gaussian, or more generally anharmonic, motions may relate to protein function. To gain a better understanding of the anharmonic fluctuations in long-time scale simulations, we previously introduced an approach called quasi-anharmonic analysis (QAA) [32]. QAA uses fourth-order statistics to describe the atomic fluctuations and summarizes the internal motions using a small number of dominant anharmonic modes. We have demonstrated this approach successfully in the con- text of both molecular recognition (ubiquitin and lysozyme) and enzyme catalysis (human cyclophilin A). For example, in ubiquitin it was observed that the anharmonic modes describe how binding regions adapt to diverse binding partners. In human cyclophilin A, QAA-derived modes identified structural and dynamical features of the transition state ensemble along the *cis/trans* isomerization reaction pathway. Finally, in the three simulations that we have examined (with different force-fields), an emergent property of QAA is that by characterizing anharmonicity in positional distributions the method discovers energetically homogeneous conformational sub-states. Note that ‘emergent’ implies that the discovered homogeneity is achieved without any prior knowledge of the internal energy of the systems.

We hypothesize that in addition to existing trajectory analysis approaches, exploiting anharmonicity to detect dynamical events and to elucidate the conformational transitions that govern these events will provide novel biophysical insights into the nature of how

internal motions affect protein function. To this end, we introduce a trajectory analysis toolbox HOST4MD: Higher-Order Statistics Toolbox for Molecular Dynamics. HOST4MD exploits fourth-order correlations in positional fluctuations and/or dihedral angle changes (1) to discover dynamical events of interest in the simulations real-time, (2) uses these events to temporally segment trajectories and (3) implements QAA to identify conformational sub-states and transitions that govern these dynamical events. We demonstrate HOST4MD on the enzyme adenylate kinase (Adk). Adk is a prototypical system for exploring protein dynamics as it undergoes dramatic conformational changes between ligand-bound and ligand-free forms [33–43]. Using open and closed (both ligand-free) Adk starting structures, we simulated the protein for a total of 0.6 μs . The use of higher-order statistics, as we will demonstrate, reveals anharmonic coupling between the LID, CORE and NMP sub-domains of Adk. In particular, HOST4MD reveals that in the open-to-closed transition of Adk, it must undergo a twisting motion resulting in a highly strained conformation of two α-helices ($\alpha 6$ and $\alpha 7$). These motions allow one to visualize distinct intermediate states that lead to energetically homogeneous conformational sub-states and also the exquisite dynamical coupling that exist between the LID and NMP sub-domains in controlling these transitions. In the closed-to-open transition, HOST4MD identifies distinct intermediate conformational sub-states that involve additional motions in the CORE sub-domain, leading to a twist in $\alpha 6$-$\alpha 7$ helices. Thus, HOST4MD allows one to map out how local motions in the $\alpha 6$ and $\alpha 7$ helices and changes in local hydrogen bond and hydrophobic interactions between the LID, NMP, and CORE sub-domains influence the global conformational changes in Adk. Taken together, HOST4MD provides a novel means to characterize Adk's conformational landscape and elucidate how its structure and dynamics control its overall function.

2. Methods

HOST4MD implements three major components as shown in Figure 1. The first component uses robust statistical approaches to find the optimal superposition for an ensemble of MD-generated conformations, a key step for any analysis of internal motions (Section 2.2). The second component uses a sliding window protocol to track any physical parameter of interest in the simulation. In this paper, we chose to track higher-order statistics (Section 2.3). We observe that the extrema in the higher-order statistics trace generated by the sliding window protocol reveal significant changes in the dynamical behavior in the protein conformation, thus providing a means to identify dynamical events of the simulation at a coarse-scale (Section 2.3). The third component implements quasi-anharmonic analysis (QAA) to identify at a finer-scale the conformational sub-states and transitions that govern the events identified at a coarser-level in the previous step (Section 2.4). Before we describe each of the components in HOST4MD, we will first explain the simulation set-up and conformational sampling for Adk.

2.1. Adk simulation set-up and conformational sampling

To demonstrate the suitability of HOST4MD for elucidating biophysically relevant anharmonic motions, we chose to investigate the model hinge protein adenylate kinase (Adk) [34, 41, 44, 45]. To ensure adequate sampling, all-atom simulations of Adk were performed using the OPLS/AA force field [46] implemented as part of the recently developed Desmond [2] program. Detailed MD simulations were performed starting with two crystal structures, 1AKE (closed conformation) [47] and 4AKE (open conformation) [48]. The two structures were first processed using the Maestro package (Schrödinger Inc.) to remove any extra chains/molecules. After adding hydrogen atoms both structures were immersed in a pre-equilibrated solvent box using the SPC water model [49]. The box size was determined such that the distance between protein surface and the boundaries was at

least 10Å. This resulted in a system with box dimensions of 66.54 × 66.54 × 66.54 $Å^3$ for 1AKE and 77.87 × 77.87 × 77.87 $Å^3$ for 4AKE.

Both systems were equilibrated using a multi-step protocol to ensure structural integrity of the protein and solvent, as described in [75]. First, solvent molecules were equilibrated using 500 steps of steepest descent. Next, conjugate gradient was used to minimize the solvent until the root mean square (RMS) of the gradients was less than 0.01 kcal/mol·Å. Next, the protein was minimized using a similar procedure to release unfavorable crystal contacts. A constant pressure simulation of 50.0 ps was performed to allow the water molecules to enter vacuous regions. The systems were then gradually heated to 300 K using the procedure outlined in [75]. Before the production runs were started, the systems were allowed to equilibrate by running a 100ps MD simulation.

Production runs were carried out using the constant number of particles, volume, and energy ensemble (NVE) at 300 K. The bond length for hydrogen atoms was constrained using SHAKE [50]. Long-range electrostatic interactions were computed using the Particle Mesh Ewald (PME) approach. Van der Waals (VDW) and electrostatic interactions were truncated at 10Å. The RESPA [51] integrator was used with a time-step of 2 fs for the production runs. The production runs lasted a total of 0.3 μs for each system and conformations were saved at regular intervals of 0.01 ns, yielding a total of 30,000 conformations for each simulation.

Computing scaled internal energy from simulations—For each conformation in the simulation, the internal energy values were evaluated using the *vrun* package within Desmond [2]. For the purposes of our simulation, only the fluctuations in the non-bonded (nb) energy terms were used. In particular, the energy evaluated E_{nb} was defined as:

$$E_{nb}=E_{elec}+E_{vdw} \quad (1)$$

where E_{elec} is the energy of electro-static interactions and E_{vdw} is the energy of all van der Waals interactions in the protein. While the overall energy of the system remains constant in an NVE simulation, the changes in the non-bonded terms varies depending on the conformation of the protein. These changes are tracked throughout the simulation and used as one of the order-parameters in our analyses, similar in spirit to the work by Kong and Karplus [52], who used it to quantify energy transduction pathways within Rhodopsin and PDZ2 domain. For visualization purposes, the distribution of energy values is normalized to have zero mean and unit variance.

2.2. Robust superposition of conformations by separating rigid from flexible regions

To visualize and understand structural variation in flexible proteins, a first step is to superimpose relevant conformers. A standard least-squares superposition estimates optimal rotation and translation parameters by minimizing the squared error between the coordinates of corresponding atoms from two selected conformations [53]. In the language of robust statistics, a least-squares solution is sensitive to gross errors [54], i.e., a large deviation arising from even a single atom can greatly distort the estimation of the transformation matrix parameters. In fact, a least-squares superposition often produces a physically inappropriate result, as it fails to distinguish structurally stable rigid residues from flexible residues, which move substantially between multiple conformations. Additionally, if a conformational change involves separate, rigid movements of one or more domains, a least-squares formulation that seeks a single set of rotational and translational parameters will fail to recover the appropriate transformation matrix [55]. While there are several approaches to address this problem [56–59] we developed a robust statistical approach to automatically determine the appropriate transformation matrix, while simultaneously being aware of

flexible and rigid parts within the protein. Our implementation is a modification of the **wRMSD** algorithm [60], where the superposition errors are Gaussian weighted so that residues that move the least have greater weighting and in turn dominate the computation of rigid body parameters.

The scale parameter of the Gaussian weighting function is set manually in wRMSD for each protein. We fix this problem by iteratively estimating the scale parameter from the empirical distribution of the superposition errors. In particular, as in [60], we will assume that superposition errors from core residues undergoing rigid motion are normally distributed, $N(0, \sigma^2)$, with zero mean and σ^2 variance. It is incorrect to estimate σ directly from an empirical distribution of the errors, which are non-negative distance values, as they can contain errors from both rigid and flexible regions whose constituent residues are yet to be determined. Instead, we first compute a more robust measure, the median of errors: median$|\{e_i\}|$ where e_i is the superposition error for atom i, with the assumption that the empirical median$|\{e_i\}|$ is approximately the same as the median of the absolute values sampled from the normal distribution $N(0,\sigma^2)$. Noting conveniently that the median of the absolute values of samples from a normal distribution is roughly 2/3 of its standard deviation σ, we set σ = 1.5 × median$|\{e_i\}|$.

In this paper, we apply the robust superposition algorithm on five different experimentally determined structures of Adk to identify a core set of rigidly moving residues (see Supporting Information (SI) Figure S1). The identified residue core is then used to align the MD ensemble under investigation.

2.3. Event detection using sliding window analysis of higher-order statistics

Measuring kurtosis for anharmonicity—As a measure of anharmonicity in protein motions, we compute kurtosis, κ, the normalized fourth central moment, defined as

$$k(q) = \frac{E\left\{(q-\mu)^4\right\}}{\sigma^4} \quad (2)$$

where $E\{q\}$, μ, and σ denote the expected value, mean, and standard deviation respectively of the random variable q. For unimodal distributions, κ is a means of quantifying their peakiness or equivalently the amount of weight in the tails. A Gaussian distribution with zero-mean and unit variance has $\kappa = 3$. A value of $\kappa > 3.0$ indicates a super-Gaussian distribution that is more peaked and heavier tailed than a Gaussian of the same variance. Conversely, a distribution with $\kappa < 3.0$, referred to as sub-Gaussian, indicates a less peaked distribution than the baseline Gaussian. We will perform a sliding window analysis of the MD trajectory to track kurtosis as described next.

Sliding window analysis—To signify time evolution of any feature of interest within the MD simulation, we will observe the MD data under an exponential window, a commonly employed practice in the signal processing literature [61]. Using an exponential window located at time t, a weight

$$W_t(k) = \alpha e^{-(t-k)/\tau}$$

is applied to the information at frame k from the past, k ≤ t (see SI Figure S2). The time constant τ for exponential weight decay is given by

$$\tau = \frac{\left(h\delta^{-1}\right)}{\log 2}$$

where h is the half-life of the envelope in nanoseconds (selected by the user), δ is the time between the consecutive frames of the MD simulation sampled for analysis (= 10 ps) and the units of τ are frames. The half-life h sets a resolution window for analyzing dynamical changes. The weights $Wt(k)$ sum to 1 by setting

$$\alpha = 1 - e^{-1/\tau}.$$

An important distinction to note here is that TimeScape uses a Gaussian window scheme to analyze the data, whereas HOST4MD uses an exponential window to give more importance to recent history than frames seen in the past.

The weights, $W_t(k)$, will be used to track the statistics of positional deviations over the duration of the simulation. The weighting scheme is a natural solution for smoothing any large fluctuations observed in time evolving properties (such as kurtosis) in the trajectories. In particular, at any given time point t they will be used to (1) find a weighted mean structure for the superposition of MD frames until time tand to (2) track any feature of interest from the simulation, including kurtosis ($\boldsymbol{\kappa}_t$).

We report kurtosis values characterizing the distribution of positional fluctuations both at a residue level and for the entire conformer at each time point t. A conformer at time t is represented by a 3N dimensional positional vector $\mathbf{x}_t$, where N is the total number of atoms under consideration. The difference in the spatial coordinates between $\mathbf{x}_t$ and the weighted mean structure of the ensemble at time t is denoted by $\delta\mathbf{x}_t$. We treat the atomic displacements in the three canonical directions to be independent and compute the kurtosis of the resulting distribution of positional deviations as indicative of the conformer-level kurtosis at time $_t$. A similar computation is repeated for positional deviations at each individual residue to obtain residue-level kurtosis.

We identify the extrema in the evolving kurtosis values from the sliding window analysis as dynamically important events. As we will show in the Results section, the sliding window analysis provides a convenient means to organize the conformational trajectory in terms of an event storyboard, where time-points denoting extrema in the conformer-level kurtosis signal important dynamical changes in Adk (Results section 3.2).

2.4. Identifying conformational sub-states using quasi-anharmonic analysis

Having established a method for identifying significant events in Adk's conformational landscape, we next elucidate the internal motions that drive the transitions from one event to the next. We will capture these internal motions with our recently developed technique called quasi-anharmonic analysis (QAA) [32]. In this framework we build a low-dimensional, linear representation of the positional deviation vectors using a set of anharmonic basis vectors:

$$\partial\mathbf{x} = \mathbf{A}\gamma \quad (3)$$

where the matrix $\boldsymbol{A}$, of size $3N \times m$ ($m \ll 3N$), is derived from an approximate diagonalization of a tensor built to hold the fourth-order statistics of positional deviations δ**x**

[62,63]. The anharmonic modes of motion $\mathbf{a}_i$ are sorted in decreasing order of their amplitudes ($||\mathbf{a}_i||$).

Each anharmonic basis vector $\mathbf{a}_i$ in the matrix $\boldsymbol{A}$ has an excitation coefficient γ_i. Similar to approaches that use principal component analysis, $\boldsymbol{A}$ fully decorrelates the ensemble of positional deviation vectors $\delta\mathbf{x}$, i.e., there are no second-order dependencies between the elements of γ. In addition, matrix $\boldsymbol{A}$ is guaranteed to reduce fourth-order dependencies. By construction, $\boldsymbol{A}$ can be non-orthogonal (unlike the normal modes from principal component analysis), meaning that excitation of an anharmonic mode $\mathbf{a}_i$ is not isolated to mode i but can potentially propagate to the other modes because of non-orthogonality. By design, QAA ignores any non-linear coupling that may exist in the fluctuations between different parts of a protein. In summary, the various steps involved in performing QAA are (see [32] for more details):

1. Identify rigid and flexible residues by applying Gaussian-weighted RMSD superposition to a set of available X-ray structures.
2. Use the rigid residues to iteratively align the MD ensemble. Find the positional deviations from the iteratively derived mean conformer.
3. Build a low-dimensional representation for the fluctuations of the backbone or C^{α} atoms using PCA. We choose a low-dimensional sub-space m that captures 95% of overall variance. This initial projection onto the top m PCA bases reduces the dimensionality of the problem from $3N$ to m and helps speed up convergence of the learning algorithm for QAA.
4. Learn the QAA matrix $\boldsymbol{A}$, sort the anharmonic modes in decreasing order of their magnitude to build a reduced-dimension, anharmonic space of coefficients: γ_1, γ_2 and γ_3.
5. Use a mixture-of-Gaussians model [64] to identify clusters in the combined space of (γ_1, γ_2, γ_3).
6. To gain additional insights into the conformational landscape, label (or color) each triplet (γ_1, γ_2, γ_3) in the anharmonic space by either experimental or computational features such as scaled internal energy, inter-domain distance, etc.

The computation of scaled internal energy has been discussed earlier. For Adk there are three inter-residue distances, representative of the separation between the LID, CORE, and NMP-binding domains, which have been used as order parameters [65]. We use the same residues: Ala55 from NMP, Ala127 from LID, and Ala194 from CORE regions of Adk, to report the inter-domain distances in each conformer.

3. Results

We use two MD simulation runs of the protein Adk in its apo form, totaling over 0.6 μs to demonstrate how HOST4MD can reveal novel features of the conformational landscape. We first present the results from the robust superposition algorithm (Sec. 3.1). We then track the conformer-level and residue-level kurtosis from the MD simulations using the sliding window protocol to identify dynamical events in the conformational space of Adk (Sec. 3.2). Finally, we use quasi-anharmonic analysis to identify at a finer-scale the conformational sub-states and transitions that govern the dynamical events identified at a coarser-scale (Sec. 3.3).

3.1. Robust-superposition of MD conformational ensembles

Given the large-scale motions observed within Adk simulations, the alignment process is particularly important since it can dramatically influence both analysis and interpretation of results. Using the robust superposition algorithm, we identified regions of Adk that are more mobile than others. We used five X-ray crystal structures of Adk from Escherichia coli, namely 1AKE [47], 4AKE [48], 1E4V [66], 1E4Y [66] and 2ECK [67]). These structures constitute a diverse set of conformations Adk can adopt as part of its functional cycle. Note that only the C^{α} coordinates were used for the robust superposition. We superposed the crystal structures with respect to the open form (4AKE) to identify rigid residues in Adk. Only residues that were commonly identified to be rigid across all five crystal structures were used to define the rigid-core of Adk.

The robust super-positioning identifies residues from $\alpha 1$, $\beta 1 - \beta 5$, $\alpha 6$, $\alpha 7$ and $\alpha 8$ which also form the CORE region of Adk [47]. A total of 41 residues are identified to be in the rigid-core and are highlighted in a blue surface representation in Figure 2A. These residues play a structural role within the Adk architecture [47, 48, 66, 67]. The LID and NMP regions, on the other hand, are identified to be very mobile and hence these regions contribute much less to the determination of the rigid-body parameters for superposition(Figure S1).

The 30,000 conformers from each of the two simulations (4AKE and 1AKE) structures were then aligned to the 41 residues identified from the previous step. The resulting alignment shows that the LID and NMP sub-domains undergo large-scale structural re-arrangements. The CORE sub-domain remains fairly stable, except for small motions along $\alpha 1$ and $\alpha 8$. We show the stabilization of σ in an iterative rigid body alignment implementation of wRMSD algorithm on the open and closed forms of adenylate kinase in SI Figure S1.

After aligning the conformers robustly to the rigid core, the histograms of the positional deviations in the open and closed form simulations of Adk reveal highly anharmonic behavior with kurtosis values (Equation 2) of 11 and 8.8 respectively (Figure 2B, 2C), thus motivating us to pursue higher-order statistics as an organization principle for trajectory analysis. We next present our results on tracking conformer-level and residue-level kurtosis using the sliding window protocol.

3.2. Online event-detection using higher-order statistics

Tracking conformer-level kurtosis to identify dynamical events in open-form (4AKE) Adk simulation—Figure 3 shows traces of conformer-level kurtosis under exponential windows of widths 1ns (black) and 5ns (blue). For illustration we show the first 100ns of the simulation (the complete trace is shown in SI Figure S3 and discussed in more detail in Section 4). We denote the extrema on these curves as dynamical events of interest. We identified 13 events at 1ns and 9 events at the 5ns resolution (Figure 3 and SI movie S1 for 4AKE). Most events are prominent at both time-scales; however, a closer examination reveals the presence of additional events in the 1ns resolution that are undetected in the 5ns window. As expected, at slower time-scales only large conformational changes and significant changes in kurtosis become evident. Furthermore, given the size of the window, kurtosis changes at 1ns resolution do not become obvious until about 2.5ns into the trajectory when the influence of an entire exponential window has accumulated.

Figure 3B shows the dynamical events organized as a storyboard. Reading from the storyboard, wecan see that the LID sub-domain closes down on the NMP/CORE sub-domains within the first 50ns of the simulation. This large-scale conformational change is controlled by several motions, as highlighted by arrows in Figure 3B. These events involve

bending of α6, followed by much larger movements in the LID sub-domain. The large motion in α6 is accompanied by complementary changes in both α3 and α4. It is also interesting to note that for the LID to close down on the CORE sub-domain (events 12-13), the intermediate conformations allow the LID to extend outward (events 2-3) as well as sideways (events 4-12). As illustrated by the SI movie S1, events identified by tracking the evolution of conformer-level kurtosis allow one to visualize large-scale displacements in addition to detailed intra-molecular events.

Tracking conformer-level kurtosis to identify dynamical events in closed-form (1AKE) Adk simulation—We performed this same analysis on the 300ns simulation of the closed conformer (1AKE) (SI Figure S4). Consistent with the analysis from the 4AKE simulations, we observe a number of distinct events that allow the closed conformation to reach the open form. However, an important distinction between the two simulations is that within the 1AKE simulation the 1ns windowing is devoid of any significant changes in the conformer-level kurtosis (average value of 4.7), but at 5ns windowing we find significant differences in conformer-level kurtosis (SI Figure S4). This indicates that dynamical changes in kurtosis in the closed conformation of Adk occur at slightly slower time-scales as compared to that of the open conformer. The events that lead Adk from the closed state to the open state are also quite distinct in comparison to the 4AKE simulation (Figure 3). Apart from the motions involved in the LID and NMP sub-domain, we observe large motions along α8 (highlighted by ellipses in SI Figure S4), which were not seen in the 4AKE simulations. Furthermore, when compared to the 4AKE simulations, the LID sub-domain does not undergo side-to-side motions; however, it does undergo a distinct twist-and-bend motion to open up the LID/NMP sub-domains.

Thus far, we have examined the conformer-level kurtosis under an exponential sliding window. As the simulations progress, tracking conformer-level kurtosis provides a means to coarsely partition the simulation trajectory into temporally distinct regions. To further aid such an analysis, we examine how changes at residue-level kurtosis can also provide insights into the dynamical behavior of Adk.

Tracking residue-level kurtosis to gain additional insights into dynamical events identified in the open-form (4AKE) Adk simulation—Figure 4A depicts the residue-level kurtosis for all C^{α} atoms in the protein at three different events (time points) in Figure 3, numbered 1 (5.29ns), 4 (20.8ns), and 13 (49.73ns), that were identified at 1ns resolution. The CORE sub-domain (1-29, 60-121, and 160-214) undergoes appreciable kurtosis variation, from an average value of 2.0 at 5.29ns to relative stability at around 2.5 at 20.8ns and 49.73ns, thus remaining sub-Gaussian in behavior. However, the kurtosis changes significantly in the LID sub-domain, undergoing a shift from an average kurtosis of 2.7 (sub-Gaussian) at 5.29ns to 4.7 (super-Gaussian) at 20.8ns. This shift corresponds to fluctuations that are less uniform, indicating a greater degree of freedom in the LID sub-domain to access extended conformations (within the 1ns analysis window). When the fluctuations become more super-Gaussian, they involve rare but large transitions stemming from displacements of the α6 helix of the protein (Figure 4B-C). This can be attributed to the dynamical motions in this localized region which tend to push the LID sub-domain away from its mean position. Thus, rare fluctuations observed at this time-point indicate a dynamical transition in Adk towards closing of the LID sub-domain on top the NMP/CORE sub-domains. Furthermore, as fluctuations shift back to being Gaussian at 49.73ns ($\kappa \approx 3.1$), the LID sub-domain completely closes down on the CORE sub-domain (SI Figure S4).

In summary, the dynamical events of interest we identified appear to involve complex motions in both open and closed simulations of Adk. We wish to probe further how internal

motions of the protein drive the event formation and release. For this, we turn to quasi-anharmonic analysis (QAA).

3.3. Quasi-anharmonic analysis of the Adk conformational landscape

3.3.1. Conformational transitions in the Adk landscape for 4AKE (open) simulation—It is instructive to run QAA on each segment of the temporally partitioned trajectory so that the internal motions driving the system from one event to the next become apparent. As discussed in the previous section, the time evolution of conformer-level kurtosis within the 4AKE simulation shows that the first50ns of the simulation is highly dynamic (Figure 3), while the last 250ns is relatively stable (SI Figure S3). It is not readily apparent what events might be driving the last 250ns of the trajectory, hence we subjected this portion of the trajectory to QAA analysis in the hope of discovering conformational sub-states relevant to function.

As outlined in Section 2.4, we learn QAA bases using a sub-space of 70 dominant collective degrees of freedom from an original 642 combining all the C^{α} atoms in Adk. After applying QAA to the resultant data, all 25,000 conformers were projected onto the top three QAA basis vectors ($\mathbf{a}_{1:3}$), illustrated in Figure 5. These three anharmonic modes separate 4AKE conformers into four different sub-states (I-IV, shown by elliptical contours, Figure 5A), where the color-scale indicates scaled internal energy. We observe that sub-states I and III share a mixture of high and low energy conformers whereas sub-states II and IV show the presence of exclusively high and low energy conformations respectively. A summary of the structural statistics gathered from each of the sub-states is summarized in Table 1.

Anharmonic coupling between NMP, LID and CORE sub-domains in 4AKE lead to distinct con-formational sub-states. As illustrated in Figure 5B, the dominant motions facilitating a sub-state I → II transition along anharmonic mode a1 involve the LID region closing down on the NMP/CORE sub-domains. This leads to a strain in the overall conformation, as this motion involves substantial bending of α6 and α7. The motions which facilitate sub-state I → III transitions along mode a2 involve rotation of the LID domain on top of the CORE sub-domain. As highlighted in Figure 5C, this rotary motion is complemented by a motion in the NMP domain (α3) (highlighted in green), which moves concurrently with the LID region. An examination of the motions from sub-state I → IV (rare state) highlights a further complex motion in the 4AKE simulation. Note that in this case, the collective motions involve a substantial twisting of the LID domain around α6 and α7 with complementary motions in the NMP domains (α3 – α5). Since this transition leads to a state with the lowest average internal energy, we believe that these motions constitute a “tension release”. The movies (M2-M7) in the SI provide a visual aid to interpret these complex motions in Adk.

To gain additional insights, we color the conformational landscape with inter-domain distances as in Figure 6. Brokaw and Chu [65] chose three representative residues, namely, Ala55 from NMP, Ala127 from LID, and Ala194 from CORE, and used their inter-residue distances as order parameters for identifying transition pathways between the open and closed states (and vice-versa). We used these same residues to characterize the extent of conformational changes that the LID, NMP and CORE sub-domains undergo in our simulations (Figure S8). Observe that in sub-state II, the LID-CORE (mean distance 30 Å; standard deviation 3Å) and NMP-LID (mean distance 35 Å; standard deviation 2.6 Å) distances are quite high. In sub-state IV, however, we observe that the LID-CORE distances are quite mixed, whereas NMP-LID distances are remarkably smaller (mean 20 Å; standard deviation 2.6Å). Note that NMP-CORE distances are not uniquely separated in any sub-state.

From the analysis of the inter-domain distances mapped onto the conformational landscape, it is quite clear that the statistical signatures of the LID, NMP, and CORE sub-domain movements are quite different from each other (Section 3.2). In the 4AKE simulation, the LID-CORE and NMP-LID distances reveal how sub-state II members are relatively 'open'; constituent conformations show a large separation between LID-CORE/NMP sub-domains (Figures 6B and 6C). With regard to global fluctuations, these motions clearly dominate the landscape. Sub-state III consists of conformers that have heterogeneous energy distributions; however, it is coherent in terms of its LID-CORE/NMP separation. Conformers in sub-state III therefore represent conformers with smaller distances between the LID-CORE sub-domains and high NMP-LID distances. This indicates that within this sub-state, Adk motions involve the closing of the LID on top of the CORE sub-domain. Sub-state IV, which has the least number of conformers (339 or 1% of the simulation) (Table 1) and thus represents a "rare" state, highlights complementary motions in the LID and NMP sub-domains where LID opens up with respect to the CORE region while the NMP sub-domain moves towards the CORE region.

3.3.2. Conformational transitions in the Adk landscape for 1AKE (closed) simulation—

We subjected the entire 1AKE trajectory to QAA analysis. As outlined in Section 2.4, we used QHA to identify a total of 70 dominant collective degrees of freedom from an original 642 combining all the C^{α} atoms in Adk. After applying QAA to the resultant data, all 30,000 conformers were projected onto the top three QAA basis vectors (a1:3), illustrated in Figure 7. The three anharmonic modes separate 1AKE conformers into four sub-states (I-IV, shown by elliptical contours). The landscape is colored by scaled internal energy values, which shows a dramatic separation of sub-states. In particular, sub-states II and IV are populated by low-energy conformers, while sub-states I and III share a mixture of high and low energy conformers. A summary of structural statistics gathered from each of the sub-states is given in Table 2.

Anharmonic conformational transitions in the closed-to-open state point to distinct intermediate states: In the 1AKE (closed) simulations, motions along the first anharmonic mode that facilitate transition from sub-state I → II involve the LID and NMP regions through a twisting motion (Figure 7B). We propose two causes of this motion: in the LID sub-domain, one can observe a substantial twist in α6 – α7, involving a rotation of the residues lining this region; in the NMP sub-domain, the motions around α3 – α4 are involved in separating the NMP and the LID sub-domains. Another prominent motion is observed in α8 (shown in blue), which moves away from α6 (shown in green). These motions also indicate a coupling between the NMP-CORE and LID regions which was not observed in the 4AKE transitions (compare Figures 5B-D and 7B-D). The motions highlighted in Figure 7C-D illustrate transitions from sub-state I → III and I → IV respectively. In these two transitions, conformational change leads to lowly populated states and involves a rearrangement of α7 and LID regions. These motions hence have a unique higher-order signature that is being captured by QAA.

We next colored the 1AKE conformational landscape with three inter-residue distance pairs as illustrated in Figure 8 (A-C). Unlike the 4AKE simulation, where only the NMP-LID distances showed good separation between sub-states, the 1AKE simulation has sub-states that have markedly different behavior with respect to the distribution of LID-CORE, NMP-LID and NMP-CORE distances. Sub-state I shows a mix of low and high distance separation between the NMP, LID, and CORE sub-domains. Sub-state II consists of conformers showing low NMP-LID (mean: 21.3 Å; standard deviation: 1.0 Å) and NMP-CORE (17.4 Å; 1.1 Å) distance distributions where as it has a medium separation in terms of the LID-CORE (26.1Å; 0.9 Å) distance. Sub-state III exhibits a clear intermediate distance separation between NMP-LID and NMP-CORE, unique from that of sub-state II; however,

in terms of the LID-CORE distance, sub-state III resembles sub-state II. Sub-state IV is quite unique in that it shares similar features to that of sub-state II in NMP-LID and NMP-CORE distances, however, with respect to the LID-CORE coordinate, it is almost exclusively populated with higher distance distributions, indicating that these conformers represent a partially open structure. Sub-state II reveals lower energy conformers that share an intermediate separation between the LID-CORE and a small separation in the NMP-CORE/LID distances.

4. Discussion

Our use of higher-order statistics in detecting molecular events of interest from long time-scale simulations raises several important questions. Specifically, in the case of Adk simulations we will discuss the additional biophysical insights gained from tracking higher-order statistics in the form of conformer-level (and residue-level) kurtosis over other traditional measures such as RMSD or knowledge-based features (LID-CORE-NMP distances) and other low-dimensional approaches (PCA).

Biophysical insights from tracking kurtosis

Inter-domain rearrangement is an intrinsic aspect of Adk and thought to facilitate catalytic competence [68]. In our simulations we observed that the LID and NMP sub-domains interact via two salt-bridges formed between residues Arg36 and Asp147 and, to a lesser extent, between Lys141 and Asp33. The primary salt-bridge involves the interaction of Arg36 $N_{\zeta 1}$ and $N_{\zeta 2}$ side-chain atoms with the $O_{\delta 1}$ and $O_{\delta 2}$ of Asp147 side-chain (SI Figure S5). Both these residues are highly conserved, evident from the sequence alignment of Adk from 94 species [69] (see SI Figure S6). In the first 45ns of the simulation, these residues do not approach the ideal distance geometry, however, their separation distance stabilizes after the salt-bridge forms at approximately 50ns. This event represents a significant conformational change as observed by the stability of conformer-level kurtosis (Figure 3 and SI Figures 2 and 4). Once the salt-bridge has stabilized (after 50 ns), the change in kurtosis is minimal for the next 250ns. However, the salt bridge breaks after 250ns, and this causes an increase again in conformer-level kurtosis, implying a dynamical transition to a subsequent state where the LID and NMP regions start to separate. The secondary salt bridge, Lys141-Asp33, is less stable, undergoing breakage and re-formation at several time-points (data not shown).

We also examined if such conformational changes become apparent by tracking commonly used metrics such as the root-mean squared deviation (RMSD) and/or inter-domain (LID-NMP, LID-CORE, CORE-NMP) distances. The Supporting Information provides an illustration of tracking the RMSD (Figure S7) and inter-domain distances (Figure S8) using the exponential windowing scheme. In addition, by projecting the conformations onto the space spanned by the NMP-CORE, NMP-LID and LID-CORE distances [65], we examined if the open (4AKE) and closed (1AKE) simulations could be separated into sub-states with distinct energetic and structural properties. As shown in Figure S9 we colored the conformations using a scaled energy value determined from each of the MD simulations. We find that for the 4AKE and 1AKE simulations, the three order parameters separate the conformers into high- and low-energy states. However, as we will discuss below, tracking motions in addition to the distances within the Adk landscape using higher-order statistics provides many additional insights into the intrinsic asymmetry in opening/ closing of the LID sub-domain in the open and closed states and how this asymmetry may affect the overall function of Adk.

Asymmetry in opening/closing of LID sub-domain controls Adk binding site motions

Our results from QAA here suggest a substantial asymmetry in the opening and closing of the LID sub-domain in Adk [43]. The higher-order statistical signatures in the two conformational landscapes suggest that the closing down of the LID on top of the CORE regions is a rare event (total population 2.61 %, Table 1, sub-state II), but its occurrence lends a bias towards higher internal energy state. This implies that the movement of the LID sub-domain towards the CORE involves more strain than the movement of the LID towards the NMP sub-domain. We can attribute this energy difference to the strain undergone by α7, which must bend substantially in order to allow the LID domain to interact with the CORE. In comparison, we observe in the 1AKE simulation that opening of the LID domain is controlled by regions that include α6, α8 and α1. Thus, QAA highlights the two complementary sets of anharmonic motions involving very different regions of the molecule that control the transitions from open-to-closed and closed-to-open states in Adk.

This asymmetry in opening/closing of the LID sub-domain in Adk imposes a mechanism by which the binding sites open/close. The closing down of the LID onto NMP involves a rotation and twist motion (Figure 5C-D), while the LID opening motion (in 1AKE simulation) is coupled to the motions of the CORE and NMP sub-domains (Figure 7B). This change in the dynamical coupling between the sub-domains is dependent on localized perturbations in the interfaces of three regions in Adk: (1) α6 and α7, (2) α8, α6 and α1 and (3) α2, α7 and α5. Experimental and computational studies already support the view that local perturbations within α6 and α7 are involved in the LID opening/closing motions [41, 45, 70]. In addition, our studies show that the perturbations in α6 and α7 also result in altering the interactions between α8 and α1 (see Figure 7B) and α2 and α5 (see Figure 7C-D). Further, the rotations of the LID domain show that it is only coupled to the complementary motions with α2 and α5 (Figure 5B-C). This implies that in addition to the perturbations in α6–7 regions, there are coordinated changes within the surrounding regions, an observation supported by NMR [70] and other computational studies [38, 41, 45]. Earlier work by Wolynes and co-workers [45,71] used coarse-grained models to probe this tension release mechanism in α6 and α7 as a possible reason for the large-scale conformational change in Adk; our simulations complement these approaches with all-atom detail as well as finer information regarding the possible conformational sub-states accessed during this transition.

It is unclear from our simulations whether the motions observed in the NMP and LID sub-domains involve partial folding/refolding [72], since the time-scales accessed by these simulations are substantially small compared to the experimental time-scales. Similarly, it must also be pointed out that although the LID closing in on the NMP domain is accessible to the substrate-free enzyme [69], the inter-domain interactions are different from that of the substrate bound structure. Another surprising aspect about our simulations was the relatively short time-period in which the LID and NMP sub-domains started to interact; within the first 50 ns the two sub-domains formed a salt-bridge that stabilized throughout the simulation (see SI Figure S3). Given the high degree of sequence conservation for the two residues involved (Arg36 and Asp147), it seems plausible that these residues may be involved in stabilization of the LID-NMP interactions as the substrates bind. Arg36 is known to play a role in the catalytic step of Adk [33, 73]; however, it is not known whether Asp147 plays a role in the catalytic step.

Comparison of QAA results with PCA based approaches

PCA based techniques such as quasi-harmonic analysis (QHA) [10] and essential dynamics (ED) [9] have also been widely used for dimensionality reduction and identify conformational motions that may be related to protein function, including that of studying

Adk motions [34, 37, 39, 42, 65]. A comparison of PCA to the results from QAA is presented in Figure S10. Projecting the conformers onto the top three PCA modes, it becomes apparent that the conformations do not separate into clear sub-states in comparison with results from QAA (Figures 5 and 7). In particular, for the 4AKE (open) simulations, it is difficult to distinguish the high/low energy conformers using the top three modes of PCA. The 1AKE simulation shows the presence of multiple states albeit sharing high and low energy conformers. In both the simulations, using the PCA based description of the ensemble; it is challenging to describe the landscape using a small number of sub-states. Furthermore, unlike QAA, due to the lack of a clear separation between the sub-states, it is also difficult to identify the directions in which the conformational coupling between the sub-domains is altered. Often, as has been empirically observed, describing functionally relevant motions in Adk requires using a linear combination of several PCA modes. In comparison, QAA provides a more direct description of these motions by tracking higher-order features and reveals how different sub-domains alter their coupling to transit from one sub-state to the other. While a more thorough comparison of the results of PCA-based techniques with that of QAA is a topic of future studies, our observations here suggest that tracking higher-order features within the landscape provide tangible and biophysically relevant insights into the overall conformational dynamics of Adk from both the 4AKE (open) and 1AKE (close) simulations.

5. Conclusion

In this paper, we presented a novel trajectory analysis approach, HOST4MD, for analyzing long time-scale MD simulations. HOST4MD uses higher-order statistics to (1) detect events that signal changes in the dynamical behavior of proteins as simulations are progressing and (2) uses QAA to reveal bio-physically relevant conformational coordinates that relate these events to the function of the protein. Such a description of the landscape at both coarse-scale (to identify events) and a fine-scale (to identify conformational transitions/ motions) can be advantageous in several situations. For example, in the context of building multi-scale models for long time-scale protein folding and/or aggregation simulations, HOST4MD can be used to first identify events that correspond to partially folded states (that correspond to structural intermediates in the folding pathway) and then used to elucidate the motions that allow the protein to access such intermediate states. The structural intermediates obtained from MD can then be used to design experiments that probe specific mutations and are also helpful for understanding how folding pathways are affected by disease related mutations. Furthermore, the anharmonic modes learned from long time-scale simulation data can also be used as generative models [74], as a means to characterize and build spatiotemporal models of the protein conformational landscape. HOST4MD will be hosted at http://code.google.com/p/host-md/ and will be freely available as a MATLAB package for download.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

AJS was a predoctoral trainee supported by NIH T32 training grant T32 EB009403 as part of the HHMI-NIBIB Interfaces Initiative. PKA acknowledges the support by ORNLs Laboratory Directed Research and Development (LDRD) funds and the computing time allocation from the National Center for Computational Sciences (BIO022). ORNL is managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract No. DEAC05-00OR22725. CSC was partially supported by grant GM086238 from NIH. CSC is grateful for the simulation time allocated via startup allocation grant on NSF Teragrid. PKA's contribution was supported by grant GM083946 from NIH.

References

[1]. Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. Nat Struct Biol. 2002; 9:646–652. [PubMed: 12198485]

[2]. Bowers KJ, Chow E, Xu H, Dror RO, Eastwood MP, Gregersen BA, Klepeis JL, Kolossvary I, Moraes MA, Sacerdoti FD, Salmon JK, Shan Y, Shaw DE. Scalable algorithms for molecular dynamics simulations on commodity clusters. SC '06: Proc 2006 ACM/IEEE Conf Supercomputing. 2006:43.

[3]. Shaw DE, Deneroff MM, Dror RO, Kuskin JS, Larson RH, Salmon JK, Young C, Batson B, Bowers KJ, Chao JC, Eastwood MP, Gagliardo J, Grossman JP, Ho CR, Ierardi DJ, Kolossváry I, Klepeis JL, Layman T, McLeavey C, Moraes MA, Mueller R, Priest EC, Shan Y, Spengler J, Theobald M, Towles B, Wang SC. Anton, a special-purpose machine for molecular dynamics simulation. SIGARCH Comp Arch News. 2007; 35:1–12.

[4]. Elsen E, Houston M, Vishal V, Darve E, Hanrahan P, Pande V. N-body simulation on gpus. SC '06: Proc 2006 ACM/IEEE Conf Supercomputing. 2006:188.

[5]. Hampton S, Agarwal PK, Alam SR, Crozier PS. Towards microsecond biological molecular dynamics simulations on hybrid processors. Intl Conf High Perf Comp Sim (HPCS). 2010:98–107.

[6]. van der Kamp MW, Schaeffer RD, Jonsson AL, Scouras AD, Simms AM, Toofanny RD, Benson NC, Anderson PC, Merkley ED, Rysavy S, Bromley D, Beck DA, Daggett V. Dynameomics: A comprehensive database of protein dynamics. Structure. 2010; 18(4):423–435. [PubMed: 20399180]

[7]. Pande VS, Baker I, Chapman J, Elmer SP, Khaliq S, Larson SM, Rhee YM, Shirts MR, Snow C, Sorin EJ, Zagrovic B. Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. Biopolymers. 2003; 68(1):91–109. [PubMed: 12579582]

[8]. van Gunsteren W, Bakowies D, Baron R, Chandrasekhar I, Christen M, Daura X, Gee P, Geerke D, Glattli A, Hunenberger P, Kastenholz M, Oostenbrink C, Schenk M, Trzesniak D, van der Vegt N, Yu H. Biomolecular modeling: Goals, problems, perspectives. Angew Chem Int Ed Engl. 2006; 45:4064–4092. [PubMed: 16761306]

[9]. Amadei A, Lissen ABM, Berendsen HJC. Essential dynamics of proteins. Proteins. 1993; 17:412–425. [PubMed: 8108382]

[10]. Karplus M, Kushick JN. Method for estimating the configurational entropy of macro-molecules. Macromolecules. 1981; 14(2):325–332.

[11]. Maisuradze GG, Liwo A, Scheraga HA. Principal component analysis for protein folding dynamics. J Mol Biol. 2009; 385(1):312–329. [PubMed: 18952103]

[12]. Ramanathan A, Agarwal PK. Evolutionarily Conserved Linkage between Enzyme Fold, Flexibility, and Catalysis. PLoS Biol. 2011; 9:e1001193. [PubMed: 22087074]

[13]. Chang CA, Chen W, Gilson MK. Ligand configurational entropy and protein binding. Proc Nat Acad Sci USA. 2007; 104(5):1534–1539. [PubMed: 17242351]

[14]. Das P, Moll M, Stamati H, Kavraki LE, Clementi C. Low-dimensional free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. Proc Nat Acad Sci USA. 2006; 103(26):9885–9890. [PubMed: 16785435]

[15]. Shao J, Tanner S, Thompson N, Cheatham T. Clustering molecular dynamics trajectories: 1. characterizing the performance of different clustering algorithms. J Chem Theory Comput. 2007; 3(6):2312–2334.

[16]. Noe F, Horenko I, Schutte C, Smith J. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. J Chem Phys. 2007; 126:155102. [PubMed: 17461666]

[17]. Muff S, Caflisch A. Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a b-sheet miniprotein. Proteins. 2008; 70:1185–1195. [PubMed: 17847092]

[18]. Deng N-J, Zheng W, Gallicchio E, Levy RM. Insights into the dynamics of hiv-1 protease: A kinetic network model constructed from atomistic simulations. J Am Chem Soc. 2011; 133(24): 9387–9394. [PubMed: 21561098]

[19]. Bowman GR, Beauchamp KA, Boxer G, Pande VS. Progress and challenges in the automated construction of markov state models for full protein systems. J Chem Phys. 2009; 131(12): 124101. [PubMed: 19791846]

[20]. Wriggers W, Stafford KA, Shan Y, Piana S, Maragakis P, Lindorff-Larsen K, Miller PJ, Gullingsrud J, Rendleman CA, Eastwood MP, Dror RO, Shaw DE. Automated event detection and activity monitoring in long molecular dynamics simulations. J Chem Theory Comput. 2009; 5(10):2595–2605.

[21]. Ramanathan A, Yoo JO, Langmead CJ. On-the-fly identification of conformational substates from molecular dynamics simulations. J Chem Theory Comput. 1011; 7(3):778–789.

[22]. Frauenfelder H, Parak F, Young RD. Conformational substates in proteins. Ann Rev Biophys Biophys Chem. 1988; 17:451–479. [PubMed: 3293595]

[23]. Northrup SH, Pearl MR, Morgan JD, McCammon JA, Karplus M. Molecular dynamics of ferrocytochrome c: magnitude and anisotropy of atomic displacements. J Mol Biol. 1981; 153:1087–1111. [PubMed: 6283085]

[24]. Doster W, Cusack S, Petry W. Dynamical transition of myoglobin revealed by inelastic neutron scattering. Nature. 1989; 337(6209):754–756. [PubMed: 2918910]

[25]. Hayward S, Kitao A, Go N. Harmonicity and anharmonicity in protein dynamics: A normal mode analysis and principal component analysis. Proteins. 1995; 23:177–186. [PubMed: 8592699]

[26]. Rasmussen BF, Stock AM, Ringe D, Petsko GA. Crystalline ribonuclease a loses function below the dynamical transition at 220 k. Nature. 1992; 357(6377):423–424. [PubMed: 1463484]

[27]. Ferrand M, Dianoux AJ, Petry W, Zaccaï G. Thermal motions and function of bacteriorhodopsin in purple membranes: effects of temperature and hydration studied by neutron scattering. Proc Nat Acad Sci USA. 1993; 90(20):9668–9672. [PubMed: 8415760]

[28]. Ichiye T, Karplus M. Anisotropy and anharmonicity of atomic fluctuations in proteins: analysis of a molecular dynamics simulation. Proteins. 1987; 2(3):236–259. [PubMed: 3447180]

[29]. Ichiye T, Karplus M. Anisotropy and anharmonicity of atomic fluctuations in proteins: implications for x-ray analysis. Biochemistry. 1988; 27(9):3487–3497. [PubMed: 3390447]

[30]. Frauenfelder H, Petsko GA, Tsernoglou D. Temperature-dependent x-ray diffraction as a probe of protein structural dynamics. Nature. 1979; 280(5723):558–563. [PubMed: 460437]

[31]. Frauenfelder H, Sligar S, Wolynes P. The energy landscapes and motions of proteins. Science. 1991; 254(5038):1598–1603. [PubMed: 1749933]

[32]. Ramanathan A, Savol AJ, Langmead CJ, Agarwal PK, Chennubhotla CS. Discovering conformational substates relevant to protein function. PLoS ONE. 2011; 6(1):e15827. [PubMed: 21297978]

[33]. Henzler-Wildman KA, Lei M, Thai V, Kerns SJ, Karplus M, Kern D. A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. Nature. 2007; 450(7171):913–916. [PubMed: 18026087]

[34]. Temiz NA, Meirovitch E, Bahar I. Escherichia coli adenylate kinase dynamics: comparison of elastic network model modes with mode-coupling (15)N-NMR relaxation data. Proteins. 2004; 57(3):468–480. [PubMed: 15382240]

[35]. Schlauderer G, Reinstein J, Schulz G. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. Structure. 1996; 4(2):147–156. [PubMed: 8805521]

[36]. Reinstein J, Vetter IR, Schlichting I, Roesch P, Wittinghofer A, Goody RS. Fluorescence and NMR investigations on the ligand binding properties of adenylate kinases. Biochemistry. 1990; 29(32):7440–7450. [PubMed: 2223775]

[37]. Pontiggia F, Zen A, Micheletti C. Small- and large-scale conformational changes of adenylate kinase: A molecular dynamics study of the subdomain motion and mechanics. Biophys J. 2008; 95(12):5901–5912. [PubMed: 18931260]

[38]. Maragakis P, Karplus M. Large amplitude conformational change in proteins explored with a plastic network model: Adenylate kinase. J Mol Biol. 2005; 352(4):807–822. [PubMed: 16139299]

[39]. Korkut A, Hendrickson W. Computation of conformational transitions in proteins by virtual atom molecular mechanics as validated in application to adenylate kinase. Proc Nat Acad Sci USA. 2009; 106(37):15673. [PubMed: 19706894]

[40]. Hanson JA, Duderstadt K, Watkins LP, Bhattacharyya S, Brokaw J, Chu J-W, Yang H. Illuminating the mechanistic roles of enzyme conformational dynamics. Proc Nat Acad Sci USA. 2007; 104(46):18055–18060. [PubMed: 17989222]

[41]. Daily MD, Phillips GN, Cui Q. Many local motions cooperate to produce the adenylate kinase conformational transition. J Mol Biol. 2010; 400(3):618–631. [PubMed: 20471396]

[42]. Cukier R. Apo adenylate kinase encodes its holo form: a principal component and varimax analysis. J Phys Chem B. 2009; 113(6):1662–1672. [PubMed: 19159290]

[43]. Beckstein O, Denning EJ, Perilla JR, Woolf TB. Zipping and unzipping of adenylate kinase: atomistic insights into the ensemble of open< — >closed transitions. J Mol Biol. 2009; 394(1): 160–176. [PubMed: 19751742]

[44]. Adén J, Wolf-Watz M. NMR identification of transient complexes critical to adenylate kinase catalysis. J Am Chem Soc. 2007; 129(45):14003–14012. [PubMed: 17935333]

[45]. Whitford PC, Gosavi S, Onuchic JN. Conformational transitions in adenylate kinase. Allosteric communication reduces misligation. J Biol Chem. 2008; 283(4):2042–2048. [PubMed: 17998210]

[46]. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL. Evaluation and reparametrization of the opls-aa force field for proteins via comparison with accurate quantum chemical calculations on peptides. J Phys Chem B. 2001; 105(28):6474–6487.

[47]. Muller CW, Schulz GE. Structure of the complex between adenylate kinase from Escherichia coli and the inhibitor ap5a refined at 1.9 resolution : A model for a catalytic transition state. J Mol Biol. 1992; 224(1):159–177. [PubMed: 1548697]

[48]. Muller C, Schlauderer G, Reinstein J, Schulz G. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. Structure. 1996; 4(2):147–156. [PubMed: 8805521]

[49]. Berweger CD, van Gunsteren WF, Müller-Plathe F. Force field parametrization by weak coupling. re-engineering spc water. Chem Phys Lett. 1995; 232(5-6):429–436.

[50]. Krautler V, van Gunsteren WF, Hünenberger P. A fast shake algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. J Comp Chem. 2001; 22(5):501–508.

[51]. Plimpton, S.; Pollock, R.; Stevens, M. Particle-mesh ewald and rrespa for parallel molecular dynamics simulations. Proc 8th SIAM Conf Parallel Processing for Scientific Computing; 1997.

[52]. Kong Y, Karplus M. Signaling pathways of pdz2 domain: a molecular dynamics interaction correlation analysis. Proteins. 2009; 74(1):145–154. [PubMed: 18618698]

[53]. Kabsch W. A solution for the best rotation to relate two sets of vectors. Acta Crystallographica. 1976; 32:922–923.

[54]. Hample, FR.; Ronchetti, EM.; Rousseeuw, PJ.; Stahel, WA. Robust Statistics: The Approach Based on Influence Functions. John Wiley and Sons; New York: 1986.

[55]. Zemla A. Lga: a method for finding 3d similarities in protein structures. Nucl Acids Res. 2003; 31(13):3370–3374. [PubMed: 12824330]

[56]. Theobald D, Wuttke D. Empirical bayes hierarchical models for regularizing maximum likelihood estimation in the matrix gaussian procrustes problem. Proc Natl Acad Sci USA. 2006; 103(49):18521–18527. [PubMed: 17130458]

[57]. Wu D, Wu Z. Superimposition of protein structures with dynamically weighted rmsd. J Mol Model. 2009; 16(2):211–222. [PubMed: 19568776]

[58]. Liu Y, Fang Y, Ramani K. Using least median of squares for structural superposition of flexible proteins. BMC Bioinformatics. 2009; 10:29. [PubMed: 19159484]

[59]. Mechelke M, Habeck M. Robust probabilistic superposition and comparison of protein structures. BMC Bioinformatics. 2010; 11(1):363. [PubMed: 20594332]

[60]. Damm KL, Carlson HA. Gaussian-weighted rmsd superposition of proteins: A structural comparison for flexibile proteins and predicted protein structures. Biophys J. 2006; 90:4558–4573. [PubMed: 16565070]

[61]. Jepson AD, Fleet DJ, El-Maraghi TF. Robust Online Appearance Models for Visual Tracking. IEEE Tran PAMI. 2003; 25(10):1296–1311.

[62]. Cardoso J-F. High-order contrasts for independent component analysis. Neural Comput. 1999; 11(1):157–192. [PubMed: 9950728]

[63]. Oja E, Hyvarinen A. Independent component analysis: algorithms and applications. Neural Network. 2000; 13(4-5):411–430.

[64]. McLachlan, G.; Basford, K. Mixture Models: Inference and Applications to Clustering. Vol. Volume 84 of Statistics: A Series of Textbooks and Monographs. Marcel Dekker; New York: 1988.

[65]. Brokaw JB, Chu J-W. On the roles of substrate binding and hinge unfolding in conformational changes of adenylate kinase. Biophys J. 2010; 99(10):3420–3429. [PubMed: 21081091]

[66]. Müller CW, Schulz GE. Crystal structures of two mutants of adenylate kinase from escherichia coli that modify the gly-loop. Proteins. 1993; 15(1):42–49. [PubMed: 8451239]

[67]. Berry MB, Bae E, Bilderback TR, Glaser M, Phillips GN. Crystal structure of adp/amp complex of escherichia coli adenylate kinase. Proteins. 2006; 62(2):555–556. [PubMed: 16302237]

[68]. Wolf-Watz M, Thai V, Henzler-Wildman K, Hadjipavlou G, Eisenmesser EZ, Kern D. Linkage between dynamics and catalysis in a thermophilic-mesophilic enzyme pair. Nat Struct Mol Biol. 2004; 11(10):945–949. [PubMed: 15334070]

[69]. Henzler-Wildman K, Kern D. Dynamic personalities of proteins. Nature. 2007; 450:964–972. [PubMed: 18075575]

[70]. Olsson U, Wolf-Watz M. Overlap between folding and functional energy landscapes for adenylate kinase conformational change. Nat Commun. 2010; 1:11. [PubMed: 20975667]

[71]. Miyashita O, Onuchic JN, Wolynes PG. Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. Proc Nat Acad Sci USA. 2003; 100(22):12570–12575. [PubMed: 14566052]

[72]. Rundqvist L, Adén J, Sparrman T, Wallgren M, Olsson U, Wolf-Watz M. Noncooperative Folding of Subdomains in Adenylate Kinase. Biochemistry. 2009; 48(9):1911–1927. [PubMed: 19219996]

[73]. Berry MB, Phillips GN Jr. Crystal structures of bacillus stearothermophilus adenylate kinase with bound ap5a, mg2+ ap5a, and mn2+ ap5a reveal an intermediate lid position and six coordinate octahedral geometry for bound mg2+ and mn2+ Proteins. 1998; 32:276–288. [PubMed: 9715904]

[74]. Savol AJ, Burger VM, Agarwal PK, Ramanathan A, Chennubhotla CS. QAARM: quasi-anharmonic autoregressive model reveals molecular recognition pathways in ubiquitin. Bioinformatics. 2011; 27:i52–i60. [PubMed: 21685101]

[75]. Ramanathan A, Agarwal PK. Computational identification of slow conformational fluctuations in proteins. J Phys Chem B. 2009; 113(52):11169–11180.

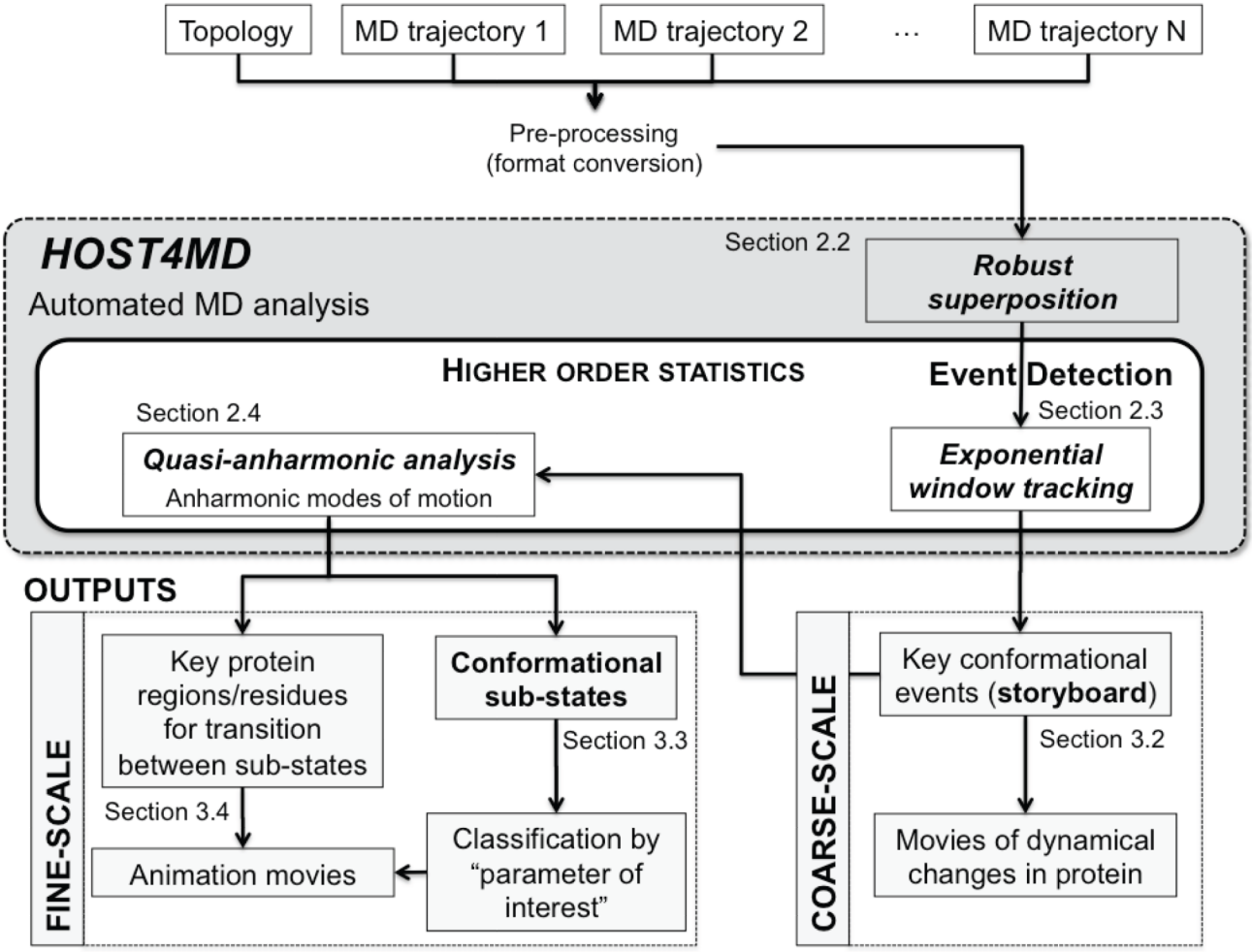

Figure 1. HOST4MD: Higher-Order Statistics Toolbox for Molecular Dynamics

HOST4MD implements three major components. The first component uses robust statistics to superpose an ensemble of conformers. The second component tracks higher-order statistics using an exponential sliding-window protocol. The third component performs quasi-anharmonic analysis. The outputs from HOST4MD include temporal segmentation of the trajectory at a coarse-scale and identification of key conformational sub-states at a fine-scale.

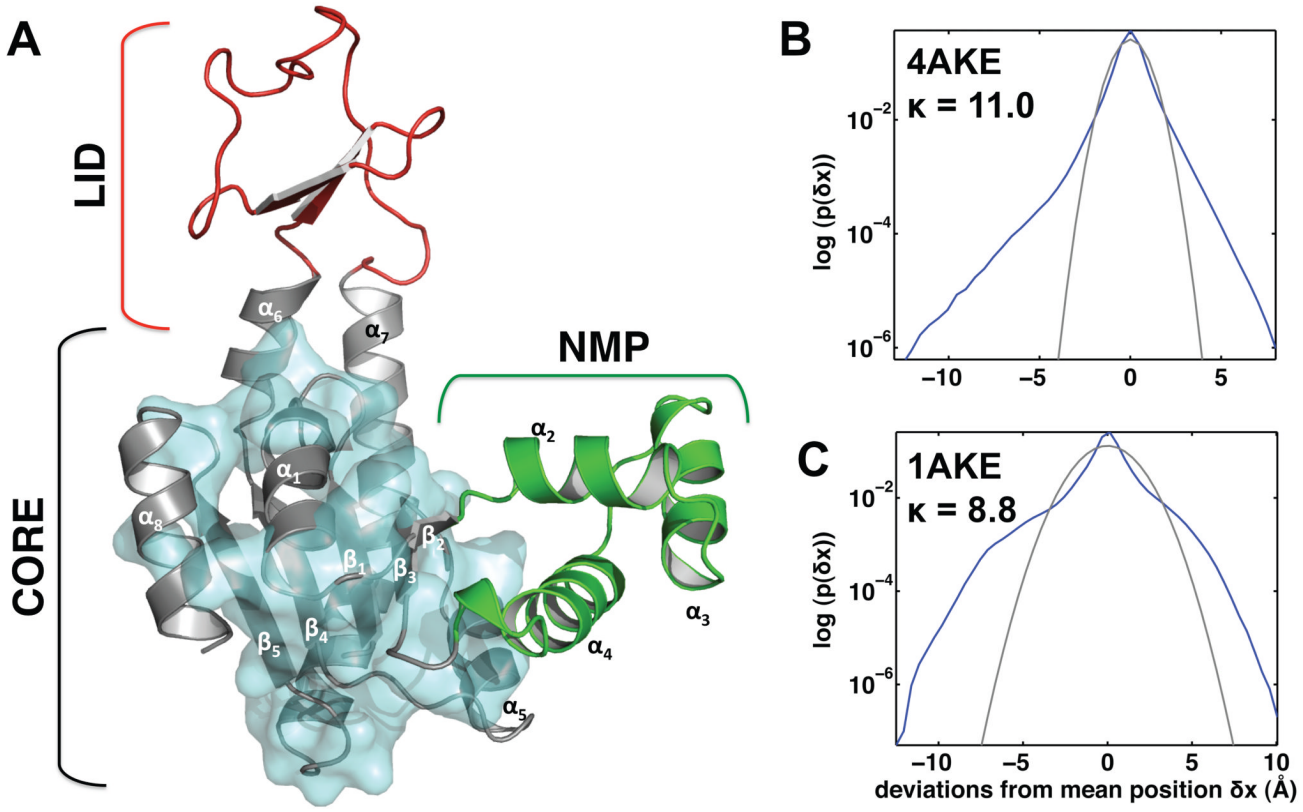

Figure 2. Robust superposition and positional deviation histograms in adenylate kinase (Adk)

(A) Rigid residues identified by the robust superposition procedure mostly come from the CORE sub-domain of Adk and are shown in cyan surface representation. The NMP and LID sub-domains are highlighted in green and red, and the constituent residues of each sub-domain are indicated [47]. Although α4 does not belong to the NMP sub-domain [47], the robust superposition procedure identifies it as being mobile and hence is weighed the least in the determination of the superposition parameters. Positional deviation histograms (blue lines) after robust superposition show non-Gaussian (anharmonic) behavior in the open (B) and close (C) simulations of Adk. The best-fitting Gaussians are shown in gray. The kurtosis values are 11 and 8.8 for the open and close simulations respectively. A color version of the figure is available online.

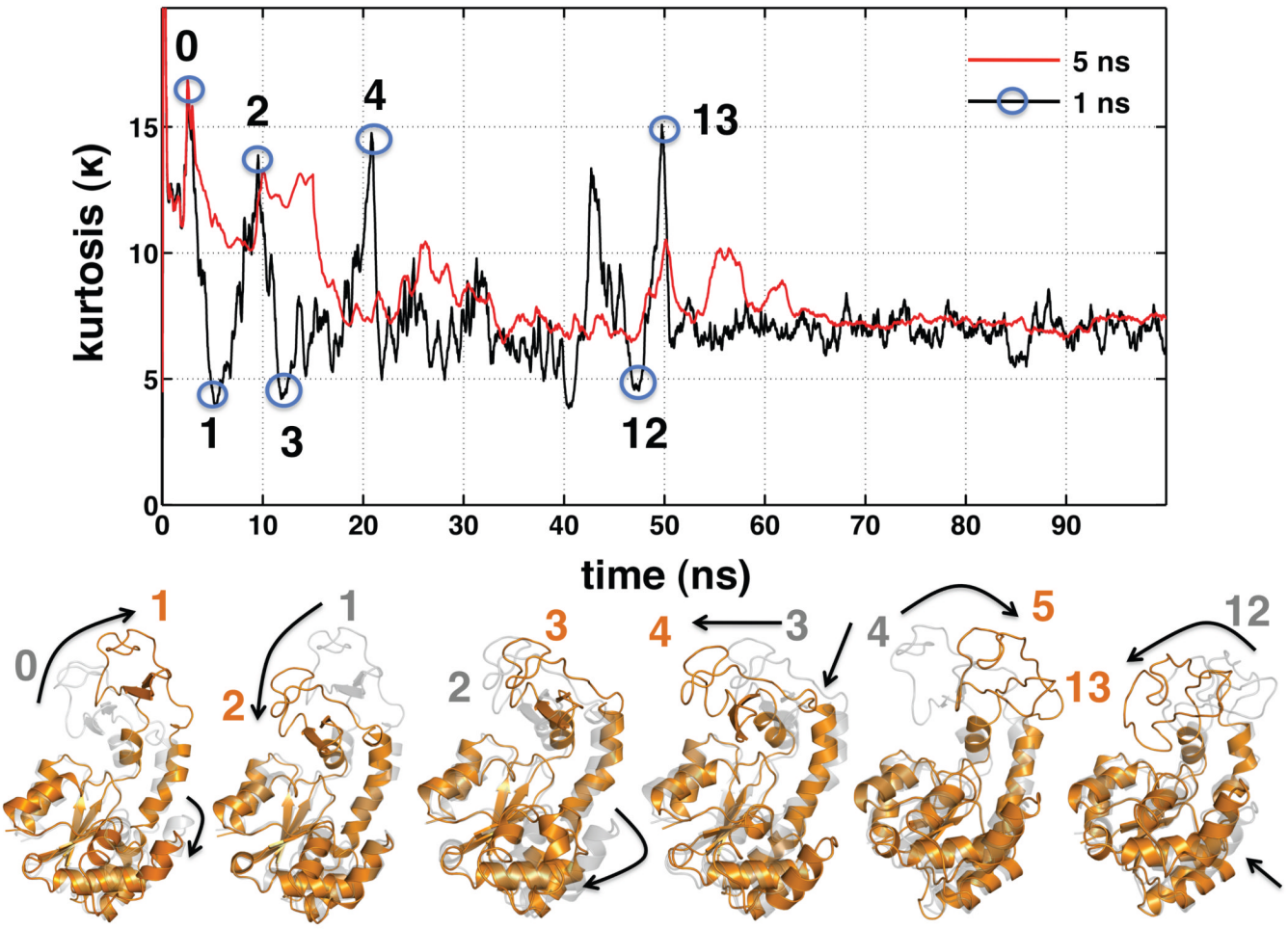

Figure 3. Online tracking of conformer-level kurtosis to identify dynamical events in 4AKE simulation

(A) The conformer-level kurtosis for the positional deviations over the C^{α} atoms is tracked with an exponential sliding window with half-lives 1ns (black) and 5ns (red). Blue circles identify a few of the extrema in the kurtosis trace. Here we show only the first 100ns of data as we observed the kurtosis value to stabilize in the last 200ns of the simulation. Kurtosis for the entire 300ns simulation is shown in SI Figure S3. (B) For illustration, we present a storyboard representation of the dynamical events identified by the blue circles. The arrows indicate the conformational changes responsible for the event formation. Observe that the LID domain undergoes large hinge bending and structural rearrangements. The CORE region (along with the rigid residues identified from Figure 2) are fairly stable throughout the events identified. A color version of the figure is available online.

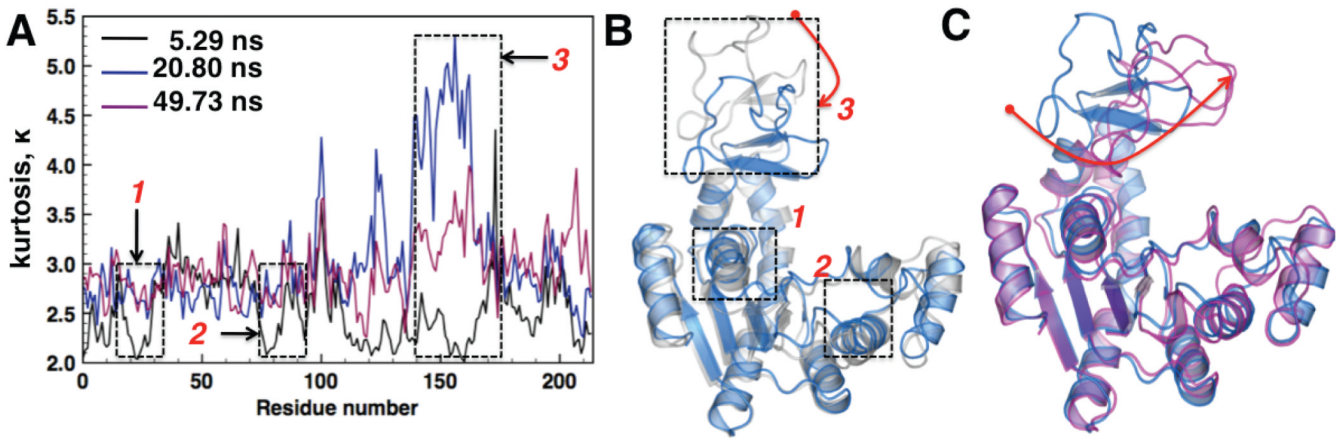

Figure 4. Comparing residue-level kurtosis at different time-points to highlight local re-arrangements in inter-residue interactions

Residue-level kurtosis measured using an exponential window with a half-life of 1ns is compared at three different time-points: 5.29ns, 20.8ns and 49.73ns corresponding to events 1, 4 and 13 respectively (in Figure 3). (B and C) Arrows and boxes indicate the three regions in Adk that show the most dynamical changes in residue-level kurtosis (also highlighted in panel A). The structures are color-coded by time-points: 5.29 ns - gray (event 1); 20.80 ns - blue (event 4) and 49.73 ns - magenta (event 13). One can distinctly note the top-down motion of the LID domain between events 1 and 4 (arrow in panel B) and a side-by-side (lateral) motion between events 4 and 13 (panel C). The residue-level kurtosis trace reveals a shift from super-Gaussian to sub-Gaussian behavior, indicating a dynamical shift in the positional fluctuations. A color version of the figure is available online.

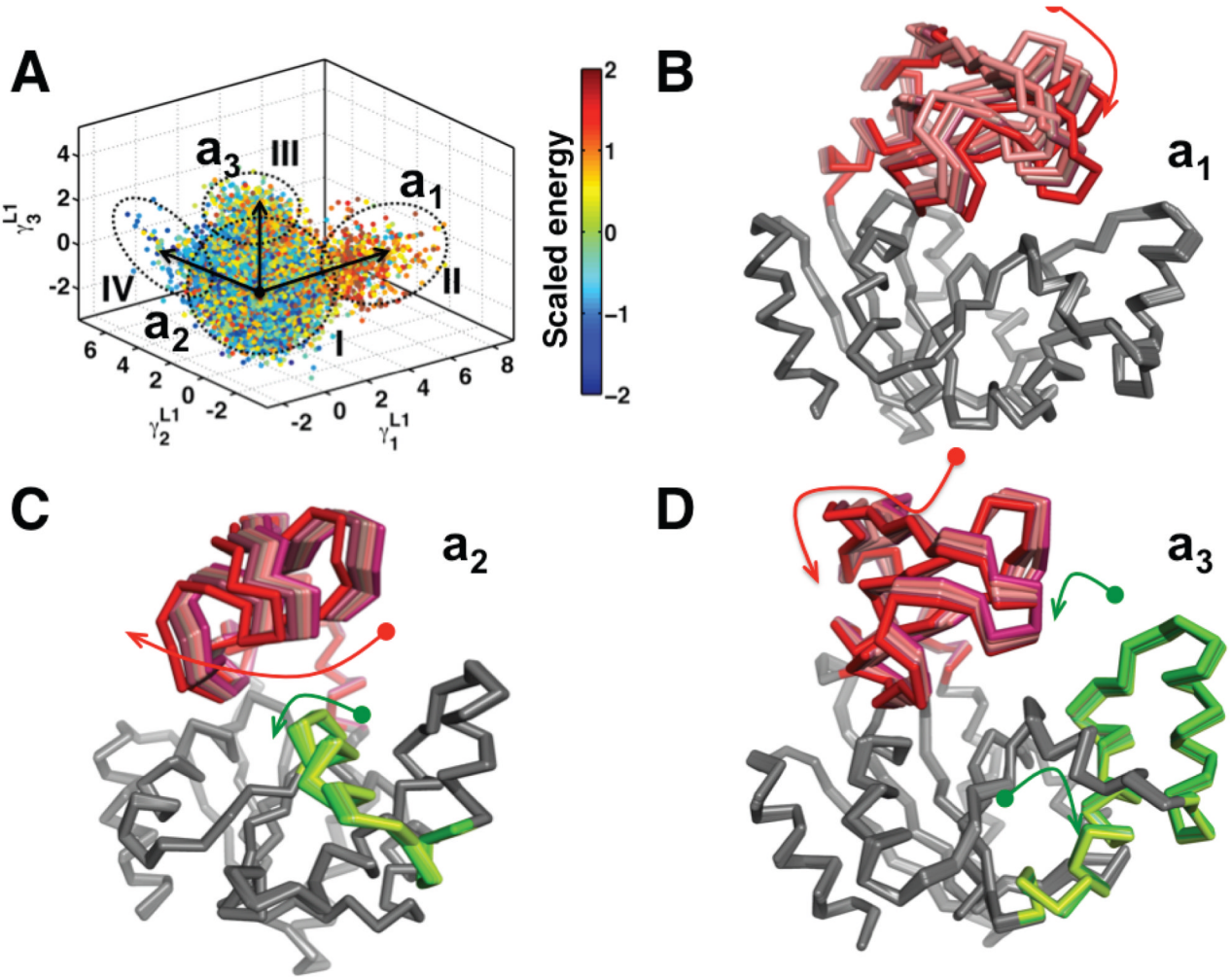

Figure 5. QAA reveals energetically homogeneous sub-states in 4AKE simulations

The last 250 ns (25,000 conformations) are projected onto the top three anharmonic modes ($\mathbf{a}_{1:3}$) from QAA. (A) Data points are painted according to scaled internal energy (sum of van der Waals and electrostatic energy terms) values. Observe that the three modes separate the landscape into energetically homogenous sub-states. (B-D) illustrate each anharmonic mode that identifies a unique motion closely associated with a transition between sub-states, represented by the arrows. Anharmonic modes also identify coupling between different regions of Adk (see Figure 6 and the main text for explanation). Movies in the Supporting Information highlight these motions. A color version of the figure is available online.

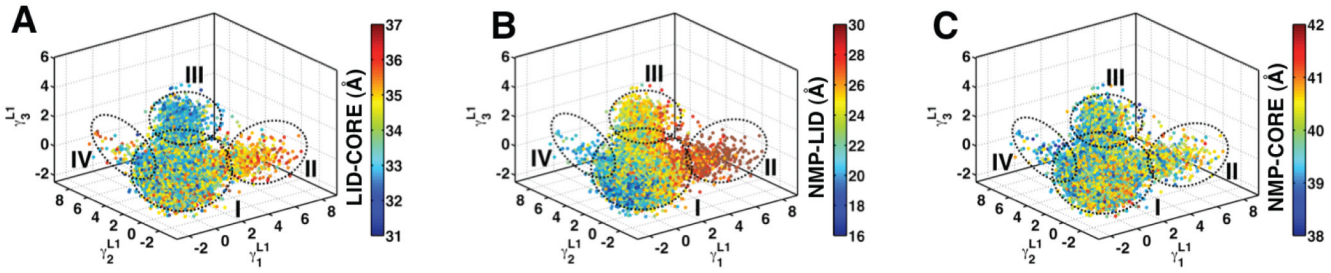

Figure 6. The intrinsic coupling between LID, CORE and NMP sub-domains in the 4AKE simulations

Same spatial projection as in Figure 5 except conformers are painted according to (A) LID-CORE distance, (B) NMP-LID distance(which is 38.8 Å in 4AKE structure) and (C) NMP-CORE distance (which is 31.3 Å in 4AKE structure) (D). See Figure S8 for a mapping of LID-CORE-NMP distances on Adk structure. Note that as Adk moves from one sub-state to the other, the coupling between the three sub-domains also changes. The color differences highlight how complex the internal motions are within the 4AKE landscape. A color version of the figure is available online.

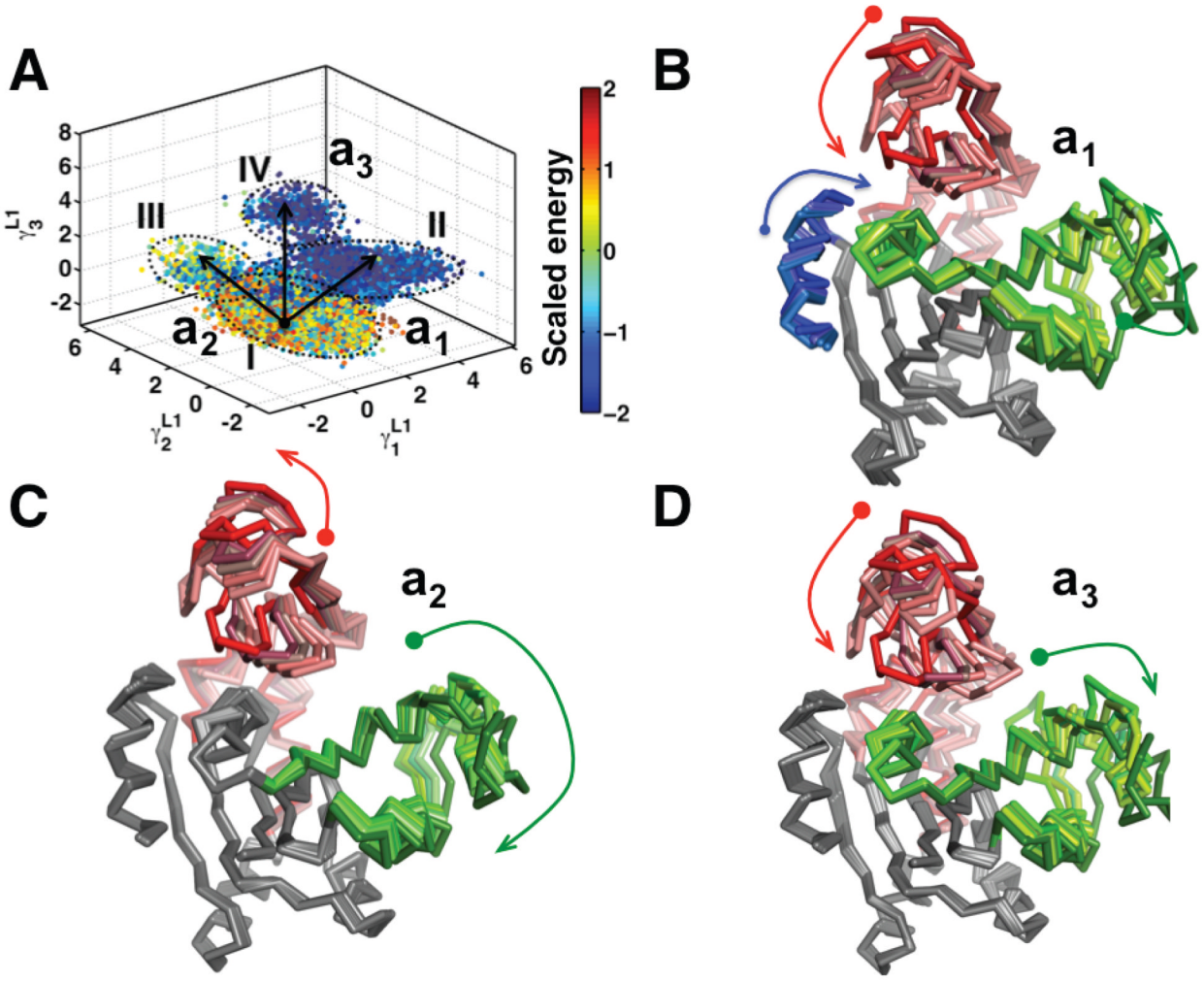

Figure 7. QAA reveals energetically homogeneous sub-states in 1AKE simulations

(A) 30,000 conformations from the 1AKE simulations projected onto the top three QAA-derived anharmonic modes ($\mathbf{a}_{1:3}$). Conformers are colored according to scaled internal energy (sum of van der Waals and electrostatic energy terms) values. Color bar wedges indicate mean scaled internal energies per cluster. (B-D) The top three anharmonic modes for 1AKE are illustrated in a cartoon like representation. The regions that move the most are highlighted in different colors; darker colors are used to indicate progress of the motion along the respective anharmonic mode. Movies in the Supporting Information highlight these motions. A color version of the figure is available online.

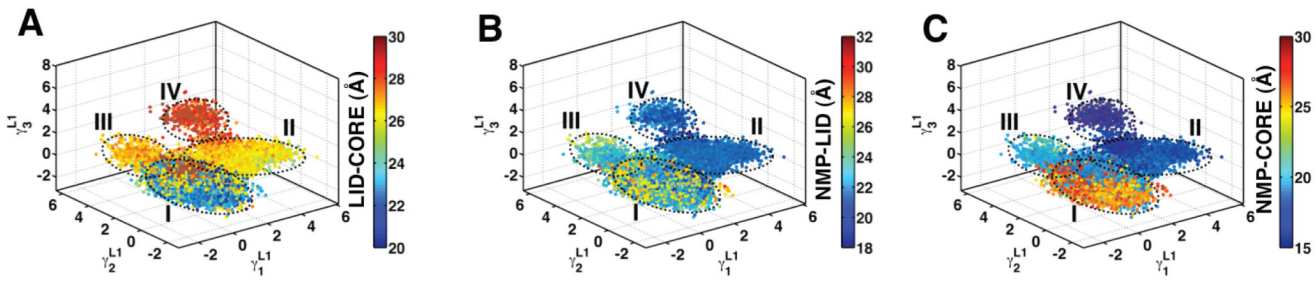

Figure 8. The intrinsic coupling between the LID, CORE and NMP sub-domains in the 1AKE simulations

Same spatial projection as in Figure 7 except conformers are painted according to (A) LID-CORE, (B) NMP-LID and (C) NMP-CORE distances. See Figure S8 for a mapping of LID-CORE-NMP distances on Adk structure. For reference the corresponding distance within the 1AKE crystal structure is indicated as a wedge on each color bar. A color version of the figure is available online.

Table 1

4AKE simulation sub-state energetics and structural features in Figure 5

Sub-states →	**I**	**II**	**III**	**IV**
Number of conformers	22487 (90%)	652 (2%)	1522 (6%)	339 (1%)
Mean scaled internal energy	−0.0320	0.5961	0.0402	0.2278
Mean LID-CORE distance (Å)	40.2	40.7	39.8	40.0
Mean LID-NMP distance (Å)	23.6	29.0	24.7	22.1
Mean NMP-CORE distance (Å)	40.7	41.0	40.3	40.3

Table 2

1AKE simulation sub-state energetics and structural features in Figure 7

Sub-states →	**I**	**II**	**III**	**IV**
Number of conformers	27314 (87%)	1311 (4%)	1588 (5%)	1087 (4%)
Mean scaled internal energy	0.1632	−1.4815	−0.3091	−1.6795
Mean LID-CORE distance (Å)	24.7	26.3	26.9	27.9
Mean LID-NMP distance (Å)	26.3	20.3	22.7	20.7
Mean NMP-CORE distance (Å)	29.9	28.0	28.5	27.7