Transition path theory analysis of c-Src kinase activation

Yilin Meng[a], Diwakar Shukla[b,c,1], Vijay S. Pande[b,c], and Benoît Roux[a,2]

[a]Department of Biochemistry and Molecular Biology, The University of Chicago, Chicago, IL 60637; [b]Department of Chemistry, Stanford University, Stanford, CA 94305; and [c]Simulation of Biological Structures NIH Center for Biomedical Computation, Stanford University, Stanford, CA 94305

Edited by Michael L. Klein, Temple University, Philadelphia, PA, and approved June 21, 2016 (received for review March 3, 2016)

Nonreceptor tyrosine kinases of the Src family are large multidomain allosteric proteins that are crucial to cellular signaling pathways. In a previous study, we generated a Markov state model (MSM) to simulate the activation of c-Src catalytic domain, used as a prototypical tyrosine kinase. The long-time kinetics of transition predicted by the MSM was in agreement with experimental observations. In the present study, we apply the framework of transition path theory (TPT) to the previously constructed MSM to characterize the main features of the activation pathway. The analysis indicates that the activating transition, in which the activation loop first opens up followed by an inward rotation of the αC-helix, takes place via a dense set of intermediate microstates distributed within a fairly broad "transition tube" in a multidimensional conformational subspace connecting the two end-point conformations. Multiple microstates with negligible equilibrium probabilities carry a large transition flux associated with the activating transition, which explains why extensive conformational sampling is necessary to accurately determine the kinetics of activation. Our results suggest that the combination of MSM with TPT provides an effective framework to represent conformational transitions in complex biomolecular systems.

transition path theory | conformational transition | Markov state models

Proteins, rather than being static molecular structures, often exhibit large-scale collective motions that are biologically essential for their function (1, 2). Dynamics and flexibility form the foundation of both the conformational selection and induced-fit mechanisms of protein–protein or protein–ligand binding (3). An energy landscape theory was proposed in the early 1990s to conceptualize protein dynamics (4) and was extensively used to characterize the protein folding problem (5) and allostery (6). According to the energy landscape theory, the complex topography of the landscape that underlies the dynamics can give rise to multiple and significantly populated conformational states (metastable states). The study of protein dynamics is not only interested in obtaining the thermodynamic properties of those metastable states but also pays significant attention to the transition pathways linking them and the kinetic information. A noteworthy example of biological significance is presented by the nonreceptor tyrosine kinases of the Src family. A well-characterized prototypical tyrosine kinase is c-Src, which plays vital roles in cellular signaling pathways (7); overactivation is key to tumorgenesis and metastasis and obesity (8). Although the configurations of the regulatory domains are important in controlling kinase activity, the conformational changes occurring within the catalytic domain itself (Fig. 1*A* and *SI Methods* for more details) is of special interest. Intramolecular motions, both at short and large length scales, control the activation of c-Src. Therefore, a better understanding of the internal motions displayed within the catalytic domain of c-Src during activation will clarify how protein kinases act as molecular switches, as well as help identify novel metastable states that are potentially suitable for the design of potent inhibitors.

Molecular dynamics (MD) simulations offer an attractive approach to examine the conformational dynamics in biological macromolecules at the atomic level. One possibility is to capture spontaneous transitions during long unbiased MD simulations (9). However, this strategy becomes inefficient for very slow transitions, because the system spends a large fraction of its time returning to regions that were previously visited due to the high barriers in the energy landscape. To accelerate the sampling of the conformational space, a biasing potential can be introduced in an MD simulation to overcome the high energy barriers; umbrella sampling (10) and metadynamics (11) are popular methods in studies of conformational transitions. An alternative strategy is to combine the information from a large number of short simulations to construct Markov state models (MSMs) representative of the process of interest (12–15). MSMs use discrete-time Markov chain to describe the conformational dynamics of proteins in terms of jumps between the microstates extracted from the simulation data, providing structural, thermodynamic, and long-timescale kinetic information (16, 17). For example, this approach was used to examine the activation of PKA regulatory subunit RIα in response to cAMP binding (18). Yang et al. (19) carried out a MSM analysis of a simplified coarse-grained model of the catalytic domain of Hck, a member of the Src family of kinases. More recently, Shukla et al. (20) constructed an MSM for the activation process in c-Src catalytic domain using ∼500 μs of aggregated sampling generated from Folding@Home.

A fundamental issue with conformational transitions occurring in complex macromolecular structures is that they are difficult to comprehend because of their inherent multidimensional nature. An elegant framework, called transition pathway theory (TPT), developed to facilitate the analysis of MSMs can help overcome this challenge (21, 22). TPT has previously been applied to the study of protein folding (17, 23) and the dynamics of HIV-1

Significance

Proteins often exhibit large-scale collective motions that are essential for biological macromolecules to perform their functions. To understand the nature of the large-scale conformational dynamics, we applied transition path theory to analyze the Markovian microstates that are obtained from extensive unrestrained molecular dynamics simulations, using the activating conformational changes in the kinase domain of c-Src nonreceptor tyrosine kinase as an example. These results elucidate the fine details of the conformational transition and offer a perspective for the interpretation of the conformational transition. Microstates that are not crucial to the thermodynamics but are important for the kinetics can be identified by transition path theory, explaining why extensive conformational sampling is needed to reproduce the accurate kinetics.

Author contributions: Y.M. and B.R. designed research; Y.M. performed research; Y.M., D.S., V.S.P., and B.R. analyzed data; and Y.M., D.S., V.S.P., and B.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]Present Address: Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana–Champaign, Urbana, IL, 61801.

[2]To whom correspondence should be addressed. Email: roux@uchicago.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1602790113/-/DCSupplemental.

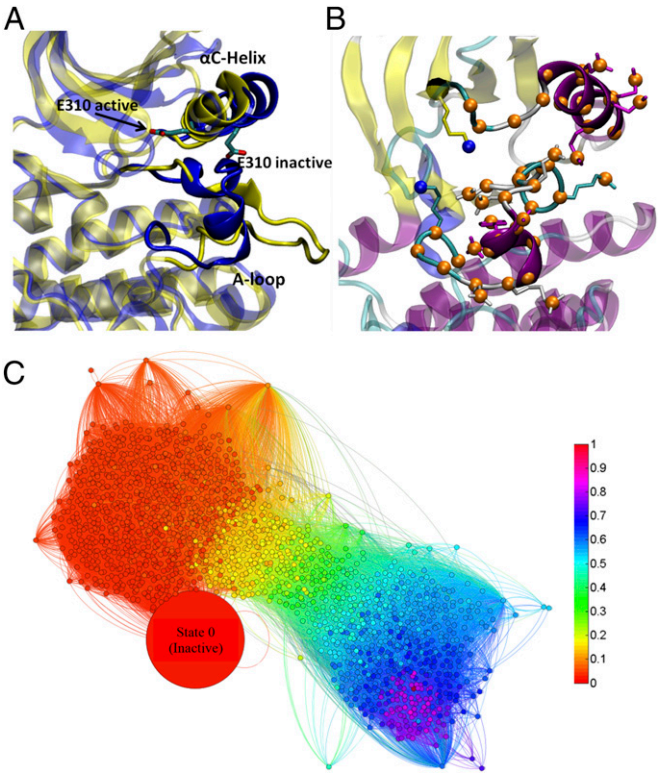

Fig. 1. (*A*) Cartoon representation of the conformational transition in c-Src catalytic domain. The inactive conformation is colored in blue, and the active conformation is colored in yellow. The main structural changes include the rotation of the αC-helix and the movement of the activation (A-) loop. E310 (chicken c-Src numbering) in the αC-helix is explicitly shown to illustrate the movement of the αC-helix in the transition. E310 is pointing outward in the inactive conformation, whereas it is pointing inward in the active-like conformation so that a catalytically important salt bridge can be formed (36). The A-loop is partially folded in the inactive kinase, but it becomes fully extended in the active conformation. (*B*) Atoms that define the space of CVs. The Cartesian coordinates of the selected atoms are used as CVs. A total of 50 atoms are selected. The carbon atoms are colored in orange, and the nitrogen atoms are colored in blue. (*C*) Network representation of the conformational transition. Each node represents a microstate from MSM, and edges represent nonzero elements of the transition probability matrix. The size of each node is proportional to the Boltzmann weight of the corresponding microstate. The largest node size is set to be 150, whereas the smallest node size is 10. Microstates are colored by their committor probability values. The program Gephi (37) was used to generate this network representation.

protease flaps (24) and has offered atomistic insights into the pathways and kinetics. The availability of a meaningful MSM analysis based on extensive MD simulation data provides a great opportunity to further examine the nature of the conformational transition pathways associated with the activation process within the c-Src kinase domain. The present study uses TPT to help elucidate the fine details of the activation transition pathway in an important signaling protein like c-Src.

Results and Discussion

Selection of the Reactant and the Product States from MSM. In a previous study, an MSM comprising 1,798 microstates with a 5-ns lag time was constructed from ∼550 μs of aggregate MD trajectories. A network (1,798 nodes and 46,089 edges) representation of the MSM is displayed in Fig. 1*C*. In this network, each node represents a microstate and the edges denote a nonzero transition probability between a pair of microstates i and j. Such a large number of nodes and edges indicates that advanced analysis scheme such as TPT is necessary to reveal the inherent features of the conformational transition. The first step in applying TPT is to define the reactant and the product states. To simplify the present analysis, a single MSM microstate was assigned to be the reactant and the product, respectively. The reactant state (inactive kinase) is selected based both on structural similarity with the crystallographic structure of the inactive c-Src kinase 2SRC (25), and on the free energy of the microstates in the MSM analysis. This leads to microstate 0: It displays the lowest free energy (Fig. S1*A*) and the rmsd of all Cα atoms relative to the kinase domain of the crystal structure 2SRC (25) is 1.0 Å. Furthermore, inspection of the structure shows that its αC-helix adopts an outward-rotated conformation and the A-loop is in the partially folded and closed conformation, which are two key features of the inactive conformation. This indicates that microstate 0 is indeed a good representation of the inactive state. Structurally, the product state is expected to resemble the kinase domain from the crystal structure of active-like c-Src 1Y57 (26). Several microstates display an rmsd for Cα atoms that is fairly small (less than 1∼2 Å). In principle, a broad basin of microstates could have been ascribed to the product state. For the sake of simplicity, a single microstate state was chosen, the one yielding the smallest possible rmsd together with a distribution of committor probabilities covering the full range [0,1] with no gaps within bins of 0.1 (see below for more discussion about committor probabilities). The microstate (1663) chosen to stand for the product state has an rmsd value of all Cα atoms of 1.5 Å relative to the kinase domain of 1Y57 (26) and yields a range of committor probabilities that is well distributed (Fig. S1*B*). The free energy difference between the reactant and the product, which can be computed with the equilibrium probabilities of the two end states, is about 3.9 kcal/mol. This value is coincidentally very close to the ΔG between the inactive and active basins after integrating the 2D potential of the mean force (2D-PMF; Fig. S2*A*). One advantage of using TPT to investigate conformational transition is that all kinetic information can be incorporated in closed form from the MSM. Using the current definition of the reactant and product states, the rate of reaction $A \rightarrow B$ (k_{AB}; see *SI Methods* for the calculation of k_{AB}) is estimated to be 1/95 μs^{-1}, based on TPT. This is consistent with the mean first passage time of 110 μs previously determined from the slowest eigenvalue from the MSM, further validating our choice of the two end states.

Analysis of All Markov Microstates. When studying conformational transitions, the reaction coordinate is a single variable that quantifies progress along a pathway. The committor probability (i.e., the probability, starting from any microstate, to first reach state A rather than state B), serves as an ideal reaction coordinate (27, 28). By constructing a histogram of the microstates based on the committor probabilities and accumulating the equilibrium probabilities in each bin, the free energy profile along the "reaction coordinate"–committor probability can be obtained. The resulting 1D free energy profile is plotted in Fig. 2*A*. One can observe the existence of multiple metastable intermediate states after lumping microstates along the reaction coordinate. The 1D free energy profile demonstrates that the basin with lowest free energy contains the inactive state of the kinase domain. The shallow free energy well centered at $q \approx 0.2$ corresponds to the intermediate state observed in the 2D-PMF calculated from replica exchange molecular dynamics/umbrella sampling calculation (the lower-right corner of Fig. S2*A*). A new free energy well centered at $q \approx 0.6$ can be found along the 1D free energy profile, indicating a second intermediate state. This intermediate state appears in the later stage of the transition and was not captured by the umbrella sampling simulations. However, this intermediate state was identified by distributed MD simulations with ∼500 μs of aggregated sampling, highlighting the importance of conformational sampling in investigating dynamical behavior of biological macromolecules. One possible reason is that umbrella sampling simulations used structural properties (linear combination of distances) as order parameters. Different regions in the 2D-PMF correspond to

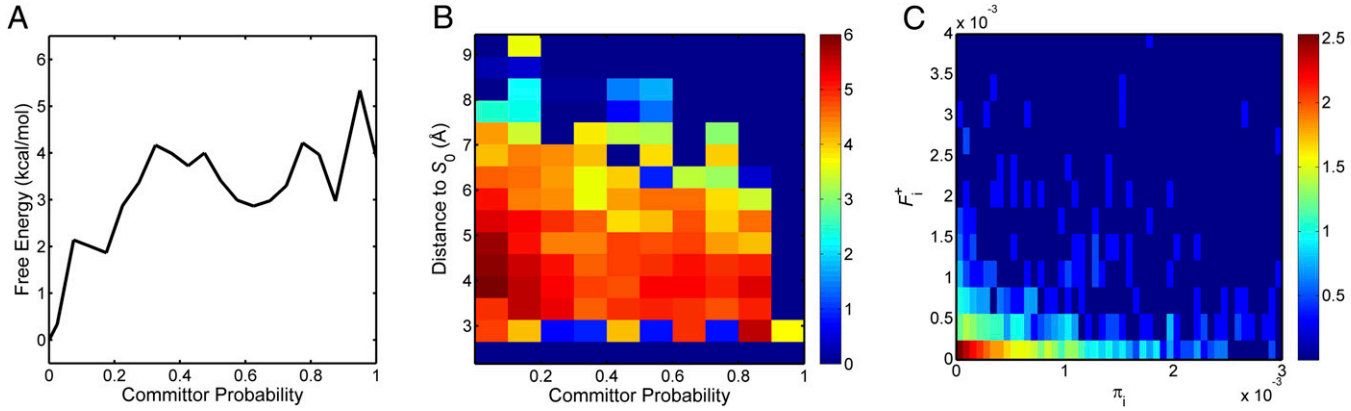

Fig. 2. TPT analysis of all MSM microstates. (*A*) Free energy profile as a function of the committor probability. (*B*) The heat map of $F^+(q_i, d_i)$. The heat map is plotted on a log10 scale of the net effective flux. Regions having F^+ to be zero are assigned a value of 10^{-7} when taking the logarithm. (*C*) Two-dimensional histogram of (π_i, F_i^+). It is zoomed-in so that nonzero counts are highlighted.

degenerate values of the committor probabilities (Fig. S2*B*). A new metastable state appears when projecting the 2D-PMF onto the reaction coordinate. A piece of evidence to support this reasoning is that visual inspection of the structural ensemble of the intermediate state centered at $q \approx 0.6$ reveals large structural flexibility, especially for the αC-helix. This can also be observed in Fig. S2*B*: Microstates having $q \approx 0.6$ are scattered in a large region of the 2D-PMF coordinates. The present analysis confirms that structural features indeed tend to be oversimplified when characterizing conformational transitions on the basis of a few chosen order parameters. Several widely used enhanced sampling strategies such as umbrella sampling and metadynamics rely on choosing a few (usually one to three) predefined structural collective variables (CVs) to coarse-grain a conformational transition (i.e., mapping a high-dimensional transition to a much lower dimensional structural CV space) and to bias the sampling in the CV space. They could potentially suffer from this oversimplified mapping and fail to capture crucial aspects of the transition. Similar issues might also be encountered with the string method, which also depends on a set of predefined CVs, although this is somewhat alleviated because the subspace of CVs is typically of a fairly large dimension.

Although the microstates were obtained from unbiased MD simulations, projection of the microstates onto 2D-PMF still helps understand features of the conformational changes and the energy landscape. According to Fig. S2*B*, opening the A-loop occurs at the very early stage of the activation, because most of the metastable states in the horizontal band have $q_i \leq 0.3$. Microstates with a committor probability around 0.5 mostly scatter in the region between the intermediate and the active states, suggesting the kinetic bottleneck (the transition states) lies between the intermediate and the active states. As expected, the free energy basin in the 2D-PMF that includes the kinase domain in an active conformation also contains metastable microstates with committor probabilities close to 1.0. However, one can still observe microstates with $q \approx 0.5$ within a basin close to the active conformation, suggesting that some slow degrees of freedom exist despite the structural similarities of those microstates. This further confirms that there is a need to go beyond a structural perspective to fully understand the inherent complexity of conformational transition in biological macromolecular systems.

The Transition Tube. When the committor probability is used as the reaction coordinate, a "reaction tube" connecting the two end states can be defined by grouping microstates according to their committor probabilities. We are particularly interested in knowing the spatial distribution of the equilibrium probability (π_i), net flux (f_{ij}^+), and the equilibrium probability of observing a reactive trajectory (π_i^R) within this transition tube because those are key quantities characterizing a conformational transition. Together, they define the shape of the transition tube. To progress in our analysis, a histogram of committor probability was constructed (bin width of 0.1). Through this discretization process, the transition tube was separated into 12 slices or cross-sections. Microstates that fall into each slice along the transition tube were then collected, and their geometric center in the CV space was computed. The geometric center corresponds to a line at the center of the tube. In the present analysis, the Cartesian coordinates of 50 atoms (Fig. 1*B*) are used to define the CV space; Euclidean distances between each microstates and the geometric center were calculated in the CV space and those Euclidean distances were normalized by the square root of the number of CV atoms (i.e., 50) to correspond more closely to the commonly used rmsd value. The effective normalized distance, referred to as d_i, will be used throughout the remainder of the analysis. Fig. S3*A* displays d_i as a function of committor probability. According to Fig. S3*A*, the microstates distributed around the corresponding geometric center display a lower bound of distance value of ~2 Å, and an upper bound that is between 6 and 9 Å. Microstates are scattered in a large region of the CV space, indicating that the transition tube is quite broad. Linear regression of the d_i values in response to committor probability resulted in a straight line with a slope of −0.38 and intercept of 4.5. The shape of the tube on the basis of the committor probability seems to resemble a spindle. The transition tube becomes wider as it leaves the reactant state and reduces its width as it approaches the product state. Eventually, all fluxes enter the one-point product state.

To probe the distribution of the net reactive flux within the transition tube, we used the row-sum of the net-flux matrix $F_i^+ = \sum f_{ij}^+$, first introduced by Vanden-Eijnden (21). The quantity F_i^+ represents the flux leaving microstate i and eventually entering the product state. In TPT, such "productive flux elements" play an important role in determining the kinetics of the conformational transition. The total flux leaving the reactant state (F_0^+) is the total transition flux. As expected, the total transition flux that leaves microstate 0 (the reactant) equals the total flux entering microstate 1663 (the product). As shown in Fig. 2*B*, large fluxes tend to populate microstates located closer to the center of the transition tube; large probability currents flow through the central region of the transition tube. To clarify the relationship between the equilibrium probabilities π_i (important for thermodynamics) and the productive flux elements F_i^+ (important for kinetics), a 2D histogram of (π_i, F_i^+) was constructed. The 2D histogram, shown in Fig. 2*C*, illustrates that states with large

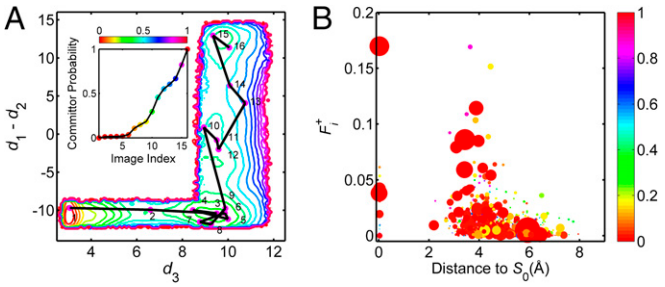

Fig. 3. Analysis of string S_0. (*A*) Projection of string S_0 onto 2D-PMF coordinates. (*Inset*) The committor probability as a function of the index of string images. (*B*) Scatter plot of the net flux of a microstate vs. its distance to S_0 in the CV space. The size of each circle is proportional to the equilibrium probability of the microstate. Microstates are colored by their committor probability values. The reactant state is excluded when making this plot because of its large Boltzmann weight and net flux.

productive flux elements F_i^+ are likely to have small equilibrium probabilities π_i, and vice versa. The TPT analysis of c-Src activation reveals that microstates that are not highly probable at equilibrium can nonetheless carry large productive fluxes to act as hubs for the transitions. Those states are important for understanding kinetics. However, microstates that possess relatively large equilibrium probability (and carry a small productive flux as shown in Fig. 2*B*) can be found far away from the geometric center of the transition tube. This simply shows that microstates that are important for thermodynamics are not necessarily important for kinetics. Transition fluxes and equilibrium probabilities need both to be taken into account to characterize conformational changes.

Reactive trajectories generated within the TPT framework, in which the system transits from the state A (inactive) to the state B (active), also provide important information on the activation mechanism. In practical applications, one is often interested in knowing the probability if a specific microstate is (or is not) along the path of a reactive trajectory (π_i^R). By definition, the reactant and the product states have $\pi_A^R = \pi_B^R = 0$. Among the remaining 1,796 microstates, state 513 has the largest probability to be part of a reactive trajectory ($\pi_{513}^R \approx 0.022$). Fig. S3*B* shows d_i with respect to q_i, with the size of each point proportional to π_i^R. Microstates having q_i between 0.1 and 0.2 are more likely to belong to a reactive trajectory than other committor probabilities. Similar to the distribution of π_i in the transition tube, microstates that are located at the outer region of the transition tube can still belong to reactive trajectories, even though those reactive trajectories do not carry a large flux.

Constructing Pathways from TPT. The string method, by which a transition pathway is discretized into a chain of states called images, is a powerful technique to study rare event in complex systems (29, 30). Once the optimized string has been obtained, thermodynamic and kinetic information can be extracted from milestoning simulations. Recent successes (31, 32) in applying the string method to study transition events prompted us to incorporate this concept in our TPT analysis of the MSM for Src kinase activation. Here we adopt the following prescription to generate path strings connecting the reactant and the product based on a steepest descent/ascent type of strategy to the net flux. Starting from the reactant (microstate 0), microstate j, which has the largest net flux f_{0j}^+, is selected. Then, microstate j is set to be the starting point and the previous criterion is applied. This leads to microstate k because f_{jk}^+ is maximal among net fluxes leaving microstate j. The process is repeated until the product state (here microstate 1663) is reached. The chain of discrete microstates can be viewed as the discretized form of a string, with increasing image index as the microstates move toward the product state. Following the above protocol, a discretized string (S_0) with 16 images can be constructed. Projection of S_0 onto the 2D-PMF is shown in Fig. 3*A*. Fig. 3*A* illustrates that S_0 mostly travels in the valley of the free energy landscape. This observation is consistent with the assumption that the optimized string corresponds to a minimum free energy pathway. Further, the intermediate state traps a large fraction of S_0 (seven images, ~44% of the images in S_0), indicating the complex nature of the transitions in the intermediate state region. Fig. 3*A* also displays that the committor probability increases monotonically along S_0. Images 11 and 12 approximately correspond to the transition state in S_0.

To analyze how the microstates distribute around S_0, a Voronoi tessellation of the CV space is used, using the 16 images of S_0 as the center of the cells. The distance between neighboring images, and the average distance from a microstate to S_0 are computed and plotted in Fig. S4*A*. The mean value of distances between adjacent cells center-to-center is 4.6 Å, with an SD of 0.6 Å. Remarkably, the sum of each segment, which is approximately the total length of the string, is ~70 Å whereas the distance between the two endpoints is only ~10 Å. The large ratio between the length of the string and the distance between two end points suggests that the optimal productive path S_0 in the CV space is highly "crumpled." The system must navigate on the rugged free energy landscape, thereby traveling a much longer total distance in the CV space than the actual distance between the end states. When viewing S_0 in the 2D-PMF coordinates, S_0 zigzags with loops in the free energy landscape. We also examined the relative position between an image (defined as the cell center) and the true geometric center of all microstates in that cell, for all cells. The distances between two types of center are also plotted in Fig. S4*A*. Comparing the average distance to S_0 (the red curve in Fig. S4*A*) and the distance between two types of centers (the blue curve in Fig. S4*A*), the latter is shorter than the former, suggesting that S_0 travels near the center of the data cloud. Moreover, the curve showing the average distance to S_0 is more or less flat, suggesting that microstates surround S_0 with equal thickness on average. A histogram of the distance from a microstate to S_0 ($d_{i\to S_0}$) is given in Fig. S4*B*. A single peak is observed and occurs at ~4.1 Å with small counts at both short and large distances.

In addition to the geometrical distribution of microstates around S_0 we are also interested in knowing the distribution of the net fluxes and equilibrium probability around S_0. To achieve this, we plot F_i^+ vs. $d_{i\to S_0}$ and use the size of a point to reflect the π_i (Fig. 3*B*). From this figure, one can see that microstates with large F_i^+ are close to S_0 (upper left of Fig. 3*B*), whereas microstates that are far away from S_0 have small F_i^+ (lower right of Fig. 3*B*). A 2D histogram of F_i^+ and $d_{i\to S_0}$ is also constructed (Fig. S5) to augment the scatter plot of F_i^+ vs. $d_{i\to S_0}$. The pattern demonstrated in Fig. 3*B* can also be observed from the 2D histogram. Furthermore, the 2D histogram shows that microstates that are 4–5.25 Å away from S_0 and have very small net fluxes are more probable than the rest. We further examine how π_i distribute around S_0. As demonstrated previously, large F_i^+ correlates with relatively small π_i, whereas large π_i correlates with small F_i^+.

Multiple strings (transition pathways) are possible for a complex system like the c-Src kinase domain. Each string carries a certain amount of transition flux. In TPT, the flux of a pathway can be calculated as the minimal flux of an edge in the pathway (17). In the case of c-Src kinase domain, the transition network contains 1,798 nodes and 46,094 edges and an edge can be used in more than one pathway. Therefore, a complete scanning of all possible pathways sorted by their fluxes is extremely time-consuming. Instead, a two-step protocol is repeated to generate a number of paths. (*i*) The steepest descent/ascent type of strategy is adopted to generate a pathway, as described above. (*ii*) Immediately following the generation of a pathway, the flux of the pathway was subtracted by all edges in that pathway. Thus, although an edge in that selected pathway could be used in other pathways, the flux is

modified every time the edge contributes to a pathway. We selected the first 200 pathways that resulted from the above procedure and used them in our analysis. Although the combined reactive flux carried by the set of 200 pathways accounts for approximately 18% of the total reactive flux from the reactant state, the set includes all of the largest flux paths. The distribution of the reactive fluxes among those 200 pathways can be found in Fig. S6. The spread of the 200 strings is similar to that of all microstates, as shown in Fig. S7, suggesting that those 200 strings are not confined in a particular region of the transition tube. The accumulated length of segments (approximated length of the string) in a string is plotted against the distance between the corresponding image and microstate 0 which is the starting image (Fig. S8) to reflect the straightness of the strings. Fig. S8 further supports that the transition tube has to travel a long distance to connect the two end points. In polymer theory, the length of a fully stretched chain L can be determined as $N{\cdot}b$, in which N is the number of segments and b is the length of each segment. Assuming the chain follows a random walk model, in which the growth of segments is independent of each other, the mean of the square of the end-to-end distance R satisfies the following relation: $\langle R^2\rangle = N \cdot b^2$. Therefore, the length of the chain can be expressed as $L^2 = N \cdot \langle R^2\rangle$. In the present study, each one of the 200 pathways can be viewed as a random chain. The end-to-end distance is the distance between the first image and the last one and is a constant (11.4 Å) because it represents the rmsd in the CV space between the reactant and the product. The average of L and N was determined to be 63.6 Å and 14, respectively, using the 200 pathways, which yield R to be ∼16 Å. One may picture as if the pathways fill the transition tube, much like a random coil in the multidimensional subspace of the CVs. Therefore, the transition pathways for the activation of c-Src correspond to random walks in the subspace of CV, confined within a reaction tube of about 4.5-Å radius rmsd. Although the activation of c-Src involves large-scale motions, the finite and relatively constant width of the reaction tube that encompasses the transition pathways does not seem to be consistent with the partial unfolding protein-quake model of conformational transition previously proposed by Miyashita et al. (6).

Milestoning Simulations Reveal Similar Structural Features. The Markovian milestoning method (33, 34) is another way to reveal both the thermodynamic and kinetic information of conformational transitions. In a Markovian milestoning simulation, partitioning the CV space into Voronoi cells is used a priori. Because the milestoning method also relies on a predefined CV subspace of low dimensionality, it too could potentially fail to capture important features of the landscape of accessible conformations.

One strategy to partition the CV space is to use a perfectly converged string representing the minimum free energy pathway. Here, the milestoning simulations started from an optimized string S_1 that was also used to seed the Folding@Home simulations. Fifty-nanosecond MD trajectories were accumulated for each Voronoi cell, and a total of 2.55 μs of MD sampling was eventually achieved. Using the 50-ns milestoning trajectories, a 1D-PMF as a function of image index (also the index of Voronoi cell centers) is computed and plotted in Fig. S9*A*, whereas the free energy profile along S_0 is shown in Fig. S9*B*. One should note here that the 1D-PMF in Fig. S9*B* is based on ∼500 μs of MD sampling. Comparing the two 1D-PMFs, one could see that the PMF from milestoning simulations yields two metastable states (one for the inactive kinase domain and the other for the active kinase domain) separated by an energy barrier. The same behavior is also seen from the TPT analysis. However, both the relative free energy between inactive and active metastable states and the free energy barrier are too high for the milestoning simulations, when quantitatively compared with the PMF yielded from MSM and TPT. To understand the difference, we recomputed the geometric center from the milestoning simulation in each Voronoi cell. A new string S_2 can be constructed by connecting adjacent geometric center. A drift from S_1 can be observed when comparing S_1 and S_2, as demonstrated in Fig. S10. Further, projecting the MSM microstates onto S_1 and S_2 also demonstrates that the microstates are closer to S_2. Our work suggests that an optimized string may seem to be a minimum free energy pathway during the optimization process of string methods (i.e., plateaus can be seen when examining the rmsd with respect to the initial and/or final strings). However, the string can still be drifting very slowly toward the true minimum free energy pathway: This drift becomes clear in a timescale of 50 ns per image in our application of the milestoning method to the Src activation. Our work indicates that obtaining a minimum free energy pathway using the string method is a slow process and requires large-scale conformational sampling. In addition, choosing/designing a metric that is able to indicate the convergence of the string method better than simply comparing the rmsd with respect to the initial and/or final strings is also an important aspect of practical application of the string method.

Conclusion

Large-scale collective motions are essential to biological macromolecules such as proteins. Independent all-atom MD simulations with explicit solvent models can be aggregated to elucidate the molecular mechanisms of such motions and to quantitatively characterize the thermodynamic and kinetic properties, based on MSMs. When the conformational transition becomes complex, more powerful theoretical frameworks such as TPT and the transition reweighted analysis method (35) can be combined with MSMs to interpret the dynamical process.

In the present study, we aimed to understand the nature of conformational transitions in c-Src tyrosine kinase, which is an example of allosteric biological macromolecules. The conformational transition that activates the c-Src kinase domain was analyzed based on aggregated data from unbiased MD simulations within a theoretical framework provided by MSM and TPT: TPT was applied to the Markov microstates from the massively distributed MD simulations. The TPT analysis indicates that the activating transition takes place via a dense set of intermediate microstates distributed within a broad and curved multidimensional "reaction tube" connecting the two end-point conformations. Large transition fluxes tend to appear near the central axis of the tube and are likely to pass through metastable microstates that are not highly populated at equilibrium. Thus, microstates with low Boltzmann weight cannot be ignored in the study of conformational transition because they may be needed to correctly represent the kinetics. Therefore, it is understandable that conformational sampling is crucial in estimating kinetic quantities accurately. However, metastable states with large equilibrium probability and with large probability to observe a reactive trajectory can be found at the outer region of the transition tube and do not necessarily contribute to the flow of transition probability current significantly. Although the pathways display considerable variability, they are nonetheless confined within a broad "reaction tube" consistent with the view that the inactive-to-active transition in c-Src kinase proceeds first by an opening of the A-loop, followed by an inward rotation of the αC-helix (19, 20).

More generally, the present analysis offers a perspective for the interpretation of a conformational transition in an important signaling allosteric protein. Our findings point out that both transition flux and equilibrium probability are essential to understand conformational changes and they need to be taken into account simultaneously. Furthermore, transition pathways populate a broad and curly reaction tube. This could explain the difficulty in representing the conformational transition via a unique and well-defined minimum free energy pathway in practice. Also, we show that metastable states with low Boltzmann weight are able to affect kinetics and should not be ignored in the study of conformational transition. Finally, our previous modeling has suggested

that compounds that stabilize the inactive and the intermediate conformations of the kinase domain might be used as c-Src inhibitors, because they can prevent activating c-Src kinase domain (20). Therefore, an understanding of the dynamic process can have a broader impact.

Methods

TPT. A brief description of TPT is given in *SI Methods*, and one could refer to Noé et al. (17) and Vanden-Eijnden (21) for more details.

Markovian Milestoning Calculation. Markovian milestoning simulations were also carried out to characterize the activating conformational transition in the catalytic domain of c-Src (see *SI Methods* for an introduction to the methodology and the computational details).

ACKNOWLEDGMENTS. Y.M. thanks Drs. Avisek Das, Mikolai Fajer, and Luca Maragliano for insightful discussions. This work was supported by National Cancer Institute, NIH Grant CAO93577 (to Y.M. and B.R.) and the Simulation of Biological Structures NIH National Center for Biomedical Computation through NIH Roadmap for Medical Research Grant U54 GM07297 (to D.S. and V.P). The computations were supported by Extreme Science and Engineering Discovery Environment Grant OCI-1053575, by NIH through resources provided by the Computation Institute and the Biological Sciences Division of the University of Chicago and Argonne National Laboratory under Grant S10 RR029030-01, and by a Biomedical Data Science Initiative Postdoctoral Fellowship from Stanford School of Medicine (D.S.).

1. Grant BJ, Gorfe AA, McCammon JA (2010) Large conformational changes in proteins: Signaling and other functions. *Curr Opin Struct Biol* 20(2):142–147.
2. Smith JC, Roux B (2013) Eppur si muove! The 2013 Nobel Prize in Chemistry. *Structure* 21(12):2102–2105.
3. Csermely P, Palotai R, Nussinov R (2010) Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem Sci* 35(10):539–546.
4. Frauenfelder H, Sligar SG, Wolynes PG (1991) The energy landscapes and motions of proteins. *Science* 254(5038):1598–1603.
5. Dill KA, MacCallum JL (2012) The protein-folding problem, 50 years on. *Science* 338(6110):1042–1046.
6. Miyashita O, Onuchic JN, Wolynes PG (2003) Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proc Natl Acad Sci USA* 100(22):12570–12575.
7. Thomas SM, Brugge JS (1997) Cellular functions regulated by Src family kinases. *Annu Rev Cell Dev Biol* 13:513–609.
8. Summy JM, Gallick GE (2003) Src family kinases in tumor progression and metastasis. *Cancer Metastasis Rev* 22(4):337–358.
9. Shaw DE, et al. (2008) Anton, a special-purpose machine for molecular dynamics simulation. *Commun ACM* 51(7):91–97.
10. Torrie GM, Valleau JP (1977) Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J Comput Phys* 23(2):187–199.
11. Laio A, Parrinello M (2002) Escaping free-energy minima. *Proc Natl Acad Sci USA* 99(20):12562–12566.
12. Chodera JD, Noé F (2014) Markov state models of biomolecular conformational dynamics. *Curr Opin Struct Biol* 25:135–144.
13. Pan AC, Roux B (2008) Building Markov state models along pathways to determine free energies and rates of transitions. *J Chem Phys* 129(6):064107.
14. Pande VS (2014) Understanding protein folding using Markov state models. *Adv Exp Med Biol* 797:101–106.
15. Noé F, Fischer S (2008) Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr Opin Struct Biol* 18(2):154–162.
16. Lane TJ, Shukla D, Beauchamp KA, Pande VS (2013) To milliseconds and beyond: challenges in the simulation of protein folding. *Curr Opin Struct Biol* 23(1):58–65.
17. Noé F, Schütte C, Vanden-Eijnden E, Reich L, Weikl TR (2009) Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc Natl Acad Sci USA* 106(45):19011–19016.
18. Boras BW, Kornev A, Taylor SS, McCulloch AD (2014) Using Markov state models to develop a mechanistic understanding of protein kinase A regulatory subunit RIα activation in response to cAMP binding. *J Biol Chem* 289(43):30040–30051.
19. Yang S, Banavali NK, Roux B (2009) Mapping the conformational transition in Src activation by cumulating the information from multiple molecular dynamics trajectories. *Proc Natl Acad Sci USA* 106(10):3776–3781.
20. Shukla D, Meng Y, Roux B, Pande VS (2014) Activation pathway of Src kinase reveals intermediate states as targets for drug design. *Nat Commun* 5:3397.
21. Vanden-Eijnden E (2014) Transition path theory. *Adv Exp Med Biol* 797:91–100.
22. E W, Vanden-Eijnden E (2010) Transition-path theory and path-finding algorithms for the study of rare events. *Annu Rev Phys Chem* 61:391–420.
23. Voelz VA, Bowman GR, Beauchamp K, Pande VS (2010) Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *J Am Chem Soc* 132(5):1526–1528.
24. Deng NJ, Zheng W, Gallicchio E, Levy RM (2011) Insights into the dynamics of HIV-1 protease: A kinetic network model constructed from atomistic simulations. *J Am Chem Soc* 133(24):9387–9394.
25. Xu W, Doshi A, Lei M, Eck MJ, Harrison SC (1999) Crystal structures of c-Src reveal features of its autoinhibitory mechanism. *Mol Cell* 3(5):629–638.
26. Cowan-Jacob SW, et al. (2005) The crystal structure of a c-Src complex in an active conformation suggests possible steps in c-Src activation. *Structure* 13(6):861–871.
27. Peters B, Trout BL (2006) Obtaining reaction coordinates by likelihood maximization. *J Chem Phys* 125(5):054108.
28. Best RB, Hummer G (2005) Reaction coordinates and rates from transition paths. *Proc Natl Acad Sci USA* 102(19):6732–6737.
29. Maragliano L, Fischer A, Vanden-Eijnden E, Ciccotti G (2006) String method in collective variables: Minimum free energy paths and isocommittor surfaces. *J Chem Phys* 125(2):24106.
30. Pan AC, Sezer D, Roux B (2008) Finding transition pathways using the string method with swarms of trajectories. *J Phys Chem B* 112(11):3432–3440.
31. Ovchinnikov V, Cecchini M, Vanden-Eijnden E, Karplus M (2011) A conformational transition in the myosin VI converter contributes to the variable step size. *Biophys J* 101(10):2436–2444.
32. Ovchinnikov V, Karplus M, Vanden-Eijnden E (2011) Free energy of conformational transition paths in biomolecules: The string method and its application to myosin VI. *J Chem Phys* 134(8):085103.
33. Maragliano L, Vanden-Eijnden E, Roux B (2009) Free energy and kinetics of conformational transitions from Voronoi tessellated milestoning with restraining potentials. *J Chem Theory Comput* 5(10):2589–2594.
34. Vanden-Eijnden E, Venturoli M (2009) Markovian milestoning with Voronoi tessellations. *J Chem Phys* 130(19):194101.
35. Wu H, Mey AS, Rosta E, Noé F (2014) Statistically optimal analysis of state-discretized trajectory data from multiple thermodynamic states. *J Chem Phys* 141(21):214106.
36. Taylor SS, Kornev AP (2011) Protein kinases: Evolution of dynamic regulatory proteins. *Trends Biochem Sci* 36(2):65–77.
37. Bastian M, Heymann S, Jacomy M (2009) Gephi: An open source software for exploring and manipulating networks. *Proceedings of the Third International AAAI Conference on Weblogs and Social Media* (Assoc for the Advancement of Artificial Intelligence, Palo Alto, CA), pp 361–362.
38. Brooks BR, et al. (2009) CHARMM: The biomolecular simulation program. *J Comput Chem* 30(10):1545–1614.
39. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79(2):926–935.
40. Zheng J, et al. (1993) 2.2 A refined crystal structure of the catalytic subunit of cAMP-dependent protein kinase complexed with MnATP and a peptide inhibitor. *Acta Crystallogr D Biol Crystallogr* 49(Pt 3):362–365.
41. Metzner P, Schutte C, Vanden-Eijnden E (2009) Transition path theory for Markov jump processes. *Multiscale Model Simul* 7(3):1192–1219.
42. Phillips JC, et al. (2005) Scalable molecular dynamics with NAMD. *J Comput Chem* 26(16):1781–1802.
43. MacKerell AD, et al. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* 102(18):3586–3616.
44. Feller SE, Zhang YH, Pastor RW, Brooks BR (1995) Constant pressure molecular dynamics simulation: The Langevin piston method. *J Chem Phys* 103(11):4613–4621.
45. Darden T, York D, Pedersen L (1993) Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J Chem Phys* 98(12):10089–10092.
46. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical integration of cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J Comput Phys* 23(3):327–341.