ACCOUNTS of chemical research

Article

pubs.acs.org/accounts

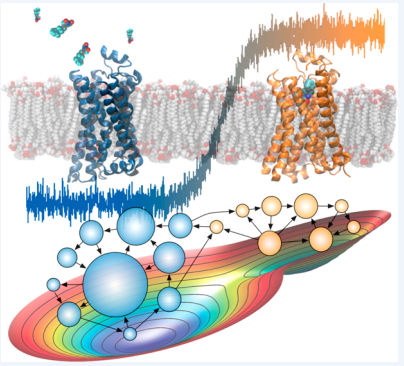
Markov State Models Provide Insights into Dynamic Modulation of Protein Function

Published as part of the Accounts of Chemical Research special issue "Protein Motion in Catalysis".

Diwakar Shukla,[†,§] Carlos X. Hernández,[‡] Jeffrey K. Weber,[†] and Vijay S. Pande*[,†,‡,§]

[†]Department of Chemistry, [‡]Biophysics Program, and [§]SIMBIOS, NIH Center for Biomedical Computation, Stanford University, Stanford, California 94305, United States

CONSPECTUS: Protein function is inextricably linked to protein dynamics. As we move from a static structural picture to a dynamic ensemble view of protein structure and function, novel computational paradigms are required for observing and understanding conformational dynamics of proteins and its functional implications. In principle, molecular dynamics simulations can provide the time evolution of atomistic models of proteins, but the long time scales associated with functional dynamics make it difficult to observe rare dynamical transitions. The issue of extracting essential functional components of protein dynamics from noisy simulation data presents another set of challenges in obtaining an unbiased understanding of protein motions. Therefore, a methodology that provides a statistical framework for efficient sampling and a human-readable view of the key aspects of functional dynamics from data analysis is required. The Markov state model (MSM), which has recently become popular worldwide for studying protein dynamics, is an example of such a framework.



In this Account, we review the use of Markov state models for efficient sampling of the hierarchy of time scales associated with protein dynamics, automatic identification of key conformational states, and the degrees of freedom associated with slow dynamical processes. Applications of MSMs for studying long time scale phenomena such as activation mechanisms of cellular signaling proteins has yielded novel insights into protein function. In particular, from MSMs built using large-scale simulations of GPCRs and kinases, we have shown that complex conformational changes in proteins can be described in terms of structural changes in key structural motifs or "molecular switches" within the protein, the transitions between functionally active and inactive states of proteins proceed via multiple pathways, and ligand or substrate binding modulates the flux through these pathways. Finally, MSMs also provide a theoretical toolbox for studying the effect of nonequilibrium perturbations on conformational dynamics. Considering that protein dynamics in vivo occur under nonequilibrium conditions, MSMs coupled with nonequilibrium statistical mechanics provide a way to connect cellular components to their functional environments. Nonequilibrium perturbations of protein folding MSMs reveal the presence of dynamically frozen glass-like states in their conformational landscape. These frozen states are also observed to be rich in β-sheets, which indicates their possible role in the nucleation of β-sheet rich aggregates such as those observed in amyloid-fibril formation. Finally, we describe how MSMs have been used to understand the dynamical behavior of intrinsically disordered proteins such as amyloid-β, human islet amyloid polypeptide, and p53. While certainly not a panacea for studying functional dynamics, MSMs provide a rigorous theoretical foundation for understanding complex entropically dominated processes and a convenient lens for viewing protein motions.

INTRODUCTION

Proteins are the main orchestrators of life, performing diverse functions in our body. For example, as you read this Account, protein (rhodopsin)[1] in your eye helps you see the text, and another protein (CAMKII)[2] controls how much information from this Account is stored in your long-term memory. The diversity in protein function originates from the different structures they adopt. Much of what we know about the protein structure−function relationship comes from the large number of protein structures obtained through X-ray crystallography. The year 2014 marks the centenary celebration of X-ray crystallography with the United Nations declaring it as the International Year of Crystallography. Over the last century, the field of X-ray crystallography (in particular macromolecular crystallography) has dramatically improved our understanding of protein function by providing much needed molecular perspectives on basic biological mechanisms.[3] Today, the Protein Data Bank contains about 100,000 crystal structures of proteins providing complex but "static" structural information. The dynamic nature of proteins has been known since the first crystal structure of myoglobin was reported in 1958.[4,5] Crystal structures represent the time and space average over a large number of molecules within a crystal (a 10 μm cubic crystal would contain ~10^{11} molecules of a 5 nm diameter protein).[6] Even such a large number of molecules in the crystal do not provide a complete dynamical picture because crystallization conditions are designed to stabilize a particular conformation of

Received: August 14, 2014
Published: January 3, 2015

DOI: 10.1021/ar5002999
Acc. Chem. Res. 2015, 48, 414−422

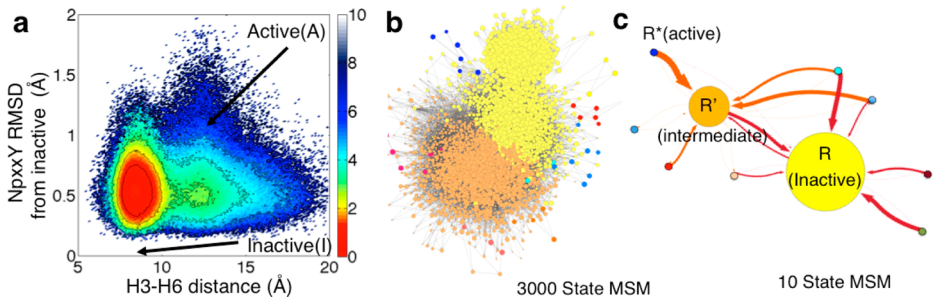


Figure 1. (a) Free energy landscape of agonist-bound GPCR (β_2-AR) using helix3−helix6 distance and the twisting of the NpxxY region in helix 7 as the order parameters. (b) Network representation of the 3000-state MSM built from the simulations of agonist-bound GPCR with each circle representing an individual conformational state. (c) Ten-state MSM built from the 3000-state MSMs using spectral clustering methods to identify kinetically relevant states. The circles in the 3000-state MSM are colored according to their membership in the coarse-grained ten-state MSM. The weight of arrow indicates the transition probability between states. Image reproduced with permission from reference 9.

the protein and are unable to represent all relevant functional states or the dynamics between them. It is now widely appreciated that the protein motions occur over a spectrum of time scales and sample a variety of conformations that can be either functionally competent (active) or incompetent (inactive). Furthermore, experimental and computational studies have shown these motions are critical for protein function.[7]

Conformational changes in proteins have been extensively studied with a particular focus on enzymes involved in a variety of disease pathways. Both experimental and computational techniques, such as nuclear magnetic resonance (NMR) spectroscopy, cryo-electron microscopy, small-angle X-ray scattering, and molecular dynamics simulations, have been successfully used to describe the protein motions and their associated time scales for the interconversion between their conformational states.[7−11] Molecular computations have a distinct advantage compared with the other methods because they provide not only the dynamic information but also the structural information in atomistic detail as a function of time.

In this Account, we provide a computational perspective on the role of protein dynamics in fundamental thermodynamics, kinetic aspects of protein function, and how ligands modulate protein dynamics. In order to demonstrate these effects, we review interesting results from the recent research performed in our laboratory that point to a promising future for this area of inquiry. These systems include protein kinases (enzymes that phosphorylate other proteins and are responsible for aberrant cellular signaling in cancer), G-protein coupled receptors (key signaling proteins that sense a wide range of extracellular signals such as hormones, drugs, photons, ions, etc.), intrinsically disordered proteins such as amyloid-β, human islet amyloid polypeptide (hIAPP), and p53 (a pivotal tumor suppressor and a target for anticancer drugs).[9,11−13] These proteins not only represent some of the most prominent targets for drug discovery but have also been studied extensively from the viewpoint of protein conformational dynamics.[14]

MARKOV STATE MODELS AND PROTEIN DYNAMICS

From a viewpoint of biology, the key questions about protein dynamics and function involve a structural definition of key conformational states of a protein, the mechanism of protein conformational change, the structure of transition states, and the height of barriers connecting these key conformations. For the physicists, the key challenge has been to reduce the complexity of the living world into simple models. The concept of an energy landscape has been widely used as a bridge connecting the disparate worlds of physics and biology. Given the appropriate degrees of freedom that describe rate-limiting protein motions (i.e., reaction coordinates), an energy landscape reduces the complexity of protein motions into a simpler human-comprehensible model. The idea of reducing the complexity of protein dynamics to a few well-chosen parameters provides significant advantages in terms of effectively eliminating the information from irrelevant protein motions and focusing on the thermodynamically and kinetically relevant motions. However, it has been shown that choosing a set of order parameters even for simple systems such as alanine dipeptide in water, pulling of DNA and RNA hairpins, etc. is nontrivial.[15] This problem arises from the large number of degrees of freedom associated with protein motions. In other words, it is difficult to construct a simple low-dimensional energy landscape for complex and entropically dominated processes such as protein conformational change without using *a priori* structural information.

Markov state models (MSMs) provide an alternate approach to these challenges by identifying the kinetically relevant states and the rates of interconversion between them. MSMs have been used extensively for modeling protein folding, and several recent studies have reported the successful use of MSMs to investigate protein conformational change.[9,11,16,17] MSMs provide a summarized view of the ensemble of spontaneous fluctuations exhibited by the protein at equilibrium by stitching together a set of individual short molecular simulation trajectories.[10,18,19] MSMs and their application to the conformational dynamics of biological systems have been the subject of several recent reviews.[8,10,20,21] Consequently, only a brief nontechnical description of MSMs is included in this Account. MSMs describe the conformational dynamics of proteins in terms of conversions between the conformational states. Similar conformations are categorized into states typically on the basis of some structural metric. The rates of interconversion between states are estimated from the simulation trajectories. Furthermore, advanced theoretical frameworks such as transition path theory (TPT)[22] could be used along with the transition probability matrix between states to identify the highest flux pathways and bottlenecks.

Finally, the number of states in a MSM can be tuned to obtain a model of desired resolution.[21] Figure 1a shows the free energy landscape associated with the conformational change in a GPCR with order parameters chosen a priori based on the available structural information. An MSM of the conformational

DOI: 10.1021/ar5002999
Acc. Chem. Res. 2015, 48, 414−422

dynamics of the same GPCR with 3000 states (Figure 1b) highlights the complexity of microstate MSMs, which could be used to observe the time evolution of any structural metric. At the same time, the high-resolution MSM could be reduced to a simple few state model, which provides the same insights about existence of an intermediate state and a rarely populated active state as the carefully chosen order parameters in Figure 1a. Therefore, we argue that MSMs represent a more natural framework for analysis of protein dynamics by embedding a high dimensional space into a more tractable representation by coarse-graining the motions in accordance with the hierarchy of time scales.

Sampling the hierarchy of time scales

Sampling the biological phenomena involving dynamics on a slow time scale (microseconds to milliseconds) has been one of the key challenges associated with molecular simulations. In principle, all conformations of the protein and their associated time scales could be obtained using large-scale molecular dynamics simulations. However, the long activation time scales push such problems out of the reach of high performance computing. One approach to surmount this challenge is to use specialized hardware and software for generating a single realization of the entire process. However, a quick look at the long trajectory of any slow time scale process would indicate that proteins spend a significant amount of simulation time fluctuating within the basins associated with long-lived metastable states and a rare thermal fluctuation leads to the barrier-crossing event. It can be shown that the waiting time for observing such rare transition is exponentially distributed. Therefore, a statistical approach that samples the rare transitions more effectively than a few long trajectories is required.

In this Account, we discuss a statistical approach for sampling the hierarchy of time scales associated with protein dynamics, which has been successfully used recently to sample conformational transitions (100 μs to millisecond time scale) associated with activation of kinases and GPCRs.[9,11] Adaptive sampling algorithms for building MSMs provide one such statistical alternative, where simulations are run in an iterative and exploratory fashion to minimize uncertainties in some property of the model.[23] The procedure for adaptive sampling comprises the iteration of three steps: running a series of short MD trajectories from initial collection of structures, building an MSM based on the aggregate data, and seeding new MD trajectories based on the sampling criterion. Weber and Pande have shown that starting new simulations from states that are least populated in the MSM provides a converged MSM with minimum number of additional simulations.[24] The convergence of MSMs at each round of sampling could be judged using relative entropy measures such as Kullback−Leibler divergence between the two MSMs, which acts as a distance metric between probability distributions in information theory. Therefore, by design, MSM-based adaptive sampling avoids well-sampled regions of the protein's conformational landscape and can effectively sample the least populated regions. Effective use of such sampling techniques in molecular dynamics studies could allow for the generation of accurate MSMs from a minimal set of short trajectories, enhancing both model accuracy and sampling efficiency.

Automatic Identification of Conformational States

Modern chemical biology and drug discovery efforts seek to develop new targets for modulating the behavior of key proteins involved in disease pathways. The issue of drug selectivity hampers the search for these novel small molecule inhibitors. Drug design for kinases, the major drug target for cancer, illustrates this problem clearly. Kinases catalyze the transfer of the γ-phosphate group from ATP to the hydroxyl group of specific serine, threonine, or tyrosine residues. The small molecule inhibitors of kinases target the highly conserved ATP-binding pocket in the inactive/active crystal structures. The inactive and active states of kinases share similar structural features due to the functional similarity between kinases and therefore provide limited selectivity. Identification of other metastable states, which are not structurally similar to inactive or active state, provides an opportunity for novel and selective drug design.[25]

Long-time scale molecular dynamics simulations could provide the sampling of the conformational landscape of proteins, but analysis of these massive simulation data sets in an unbiased manner presents a major challenge.[10] In other words, how do we turn these massive data sets into scientific insights about protein dynamics? Traditional analysis approaches involve watching movies of protein dynamics and inspecting the time evolution of order parameters identified mainly from differences in available crystal structures. Machine learning approaches and quantitative methods like principal component analysis (PCA)[26] have been used to identify key conformational motions, but these methods fail to exploit the kinetic information embedded in the MD data sets.[27] The big challenge in MSM construction involves obtaining an appropriate definition of the state space to discretize the conformational space into discrete states.

Recently, McGibbon et al.[28] have reported the use of hidden Markov models (HMMs) toward protein dynamics. These models are built on the idea that the complex protein dynamics in a large number of degrees of freedom could be reduced to a single time series representing dynamics within the set of few conformational states. The states are represented as emission (multivariate normal functions) distributions in the space of a large set of selected degrees of freedom such as distances between all α-carbons, distance of center of mass of protein side chains from their position in the reference structure, etc. In HMMs, the states do not represent a discrete partition of the conformational space of the system but provide a probabilistic estimate of observing a particular conformation as part of the each HMM state. The HMM approach simultaneously optimizes both the state decomposition (mean and distribution of the emission distribution corresponding to each state) along with the transition probabilities among states, thus providing a procedure for optimal construction of the MSMs. This is the main advantage of using HMMs over regular MSMs. The reversible hidden Markov models have been successfully applied for understanding the activation mechanism of c-src kinase (Figure 2).[28] The model not only correctly identifies the inactive and active state of the src kinase also identifies the intermediate state along the activation pathway. Furthermore, the unfolding of activation loop (red) and switching of the electrostatic network involving Lys295, Glu310 and Arg409 are the hallmark of the src kinase activation mechanism; the model identifies these metrics without any *a priori* structural information about the active state of kinase. Similarly, a strategy called time independent component analysis (tICA) has been recently developed to identify metrics that best differentiate slow modes of protein motion.[29,27] The method uses independent component analysis (ICA) to identify the

DOI: 10.1021/ar5002999
Acc. Chem. Res. 2015, 48, 414−422

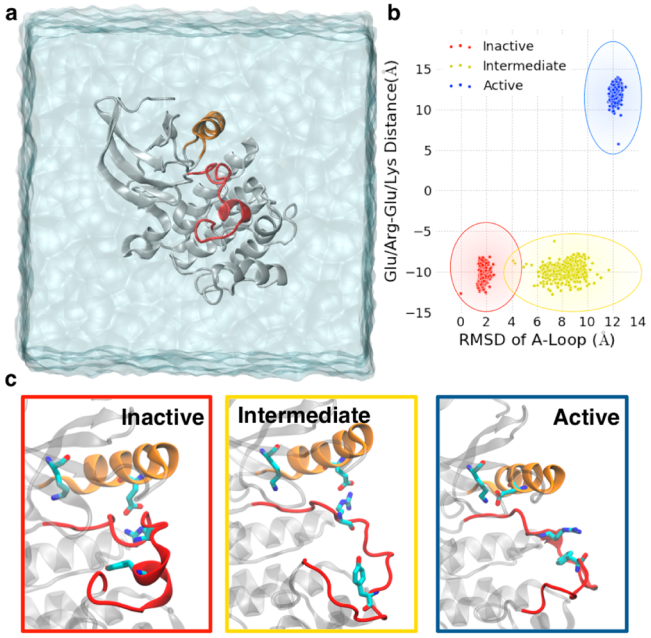


Figure 2. Three state hidden Markov model identifies the key conformational states along the activation pathway of c-src kinase. (A) Structure of the c-src kinase. (B) The projection of HMM states onto two degrees of freedom representing the RMSD of activation loop from inactive crystal structure (2SRC) and switching of the electrostatic network, respectively. (C) Snapshots of the three HMM states showing atomistic details of the activation pathway. Image reproduced with permission from reference 28.

metrics, and the slowest decorrelating principal components are then used for partitioning the conformational space of a protein. The methodology has been successfully used to study conformational transitions in peptoids[30] and folding dynamics of NTL9.[27]

In the above sections, we reviewed how Markov state models present a natural framework for sampling and analysis of protein dynamics. Application of these models to challenging scientific problems has yielded novel insights into mechanism of protein conformational change and function. In the subsequent sections, we review the key scientific insights obtained from this unbiased analysis of large protein dynamics data sets.

MODULATION OF PROTEIN FUNCTION

Molecular Switches

Identification of functionally relevant substructures called "molecular switches" within the protein is key to understanding their activation mechanism. Crystallographic data along with MD simulations and other theoretical techniques have provided a mechanism of activation of key cellular signaling proteins in terms of the conformational switching of individual molecular switches.[31,32] Recently, Kohlhoff et al. used MD simulation on Google Exacycle[33] to sample the conformational landscape of β_2-AR (β_2-adrenergic receptor, a GPCR that interacts with hormones or neurotransmitters such as adrenaline).[9] One of the key results of this study is that molecular switches related to the activation mechanism of β_2-AR can be identified by *in silico* methods (Figure 3a). These structural elements of β_2-AR have been identified by several decades of experimental research on GPCR activation. This study reports a new paradigm where large-scale simulations of GPCR could help identify these key residues for another GPCR in a matter of several months. Figure 3b shows the long-time scale behavior of the molecular switches in β_2-AR. The long time scale trajectories are generated using a kinetic Monte Carlo scheme, which provides a series of states visited over time starting from a particular MSM state. The next state in the series is picked depending upon the probability of transition from the current state to all other states. The jump between states corresponds to an increment of τ (lag time of the model) in real time. The 150 μs trajectories of β2-AR (Figure 3b) show the toggling of individual molecular switches required for the GPCR activation.

Similarly, in our recent work on the activation mechanism of src-kinase, we have not only identified the key "molecular switches" but also show that the key metastable states of src

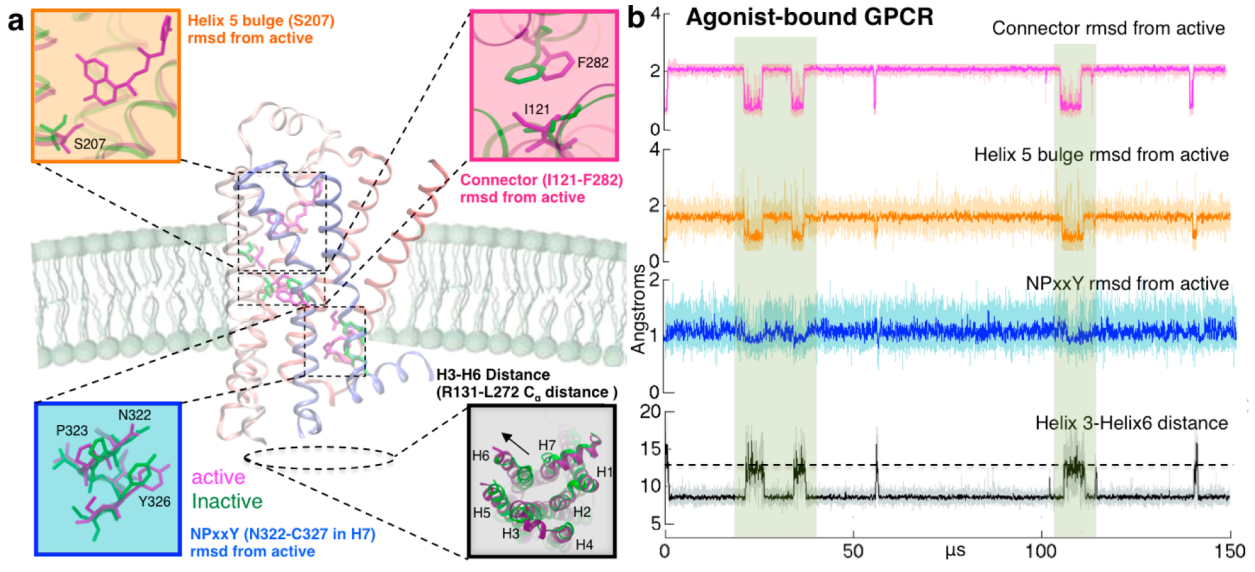


Figure 3. (a) Molecular switches involved in β_2-AR activation. (b) MSM trajectory showing the toggling of individual molecular switches during activation. Stitching together individual short MD trajectories using a kinetic Monte Carlo scheme on the MSM transition probability matrix generated these long trajectories. Image reproduced with permission from reference 9.

DOI: 10.1021/ar5002999
Acc. Chem. Res. 2015, 48, 414−422

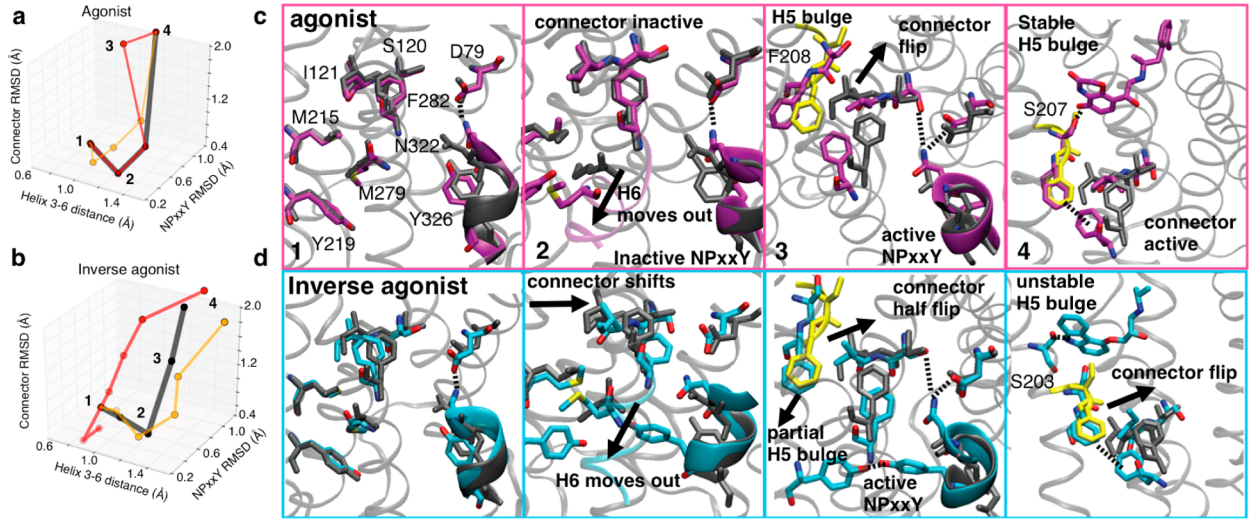


Figure 4. Modulation of GPCR-activation pathways by ligands. Activation pathways adopted in the presence of an agonist (A) and inverse-agonist (B). The corresponding structural changes along the highest flux activation pathway in the presence of an agonist (C) and inverse-agonist (D). The pathways are obtained using transition path theory on MSM transition probability matrix. Image reproduced with permission from reference 9.

kinase involve "toggling" of the individual molecular switches.[11] The inactive and intermediate states (Figure 2) differ in terms of the conformation of the activation loop, which is folded in the inactive state and unfolded in the intermediate state. The intermediate and the active states differ in the state of the electrostatic switch, which involves E310−R409 H-bond in the intermediate state and K295−E310 H-bond in the active state. The results show that MSMs of MD data sets not only can capture the novel intermediate states of proteins but also can automatically identify the key structural features involved in the activation of proteins.

Modulation of Protein Function via Multiple Pathways

Minor changes in the molecular structure of a drug or ligand are sufficient for biasing the protein function. For example, monoamines in chocolate and the psychedelic drug LSD bind the same GPCR but induce different physiological responses through G-protein and β-arrestin dependent signaling pathways, respectively. The ligand bias theory suggests that differences in the chemical structure of drugs change the protein's conformational ensemble. However, the exact nature of these structural changes has not been elucidated. Our recent work has shown that GPCRs and kinases exist in multiple conformational states, with active and inactive states connected via multiple pathways. Furthermore, we have found that ligands modulate the protein function by redistributing flux along multiple pathways. Figure 4a,b shows the projection of highest flux pathways on the three molecular switches (Figure 3) that control GPCR activation. Figure 4c,d shows how ligands alter the sequence of events along the activation pathway by allosterically modifying the conformational preferences of molecular switches. Similarly, for kinase activation, multiple pathways connect the two functional states indicating a generic mechanism where different ligands modulate the stability of different states and thereby influence the overall function.

Nonequilibrium Perturbations

Comprehending cellular dynamics on the nanoscale represents the next great frontier in biophysics. A complete "control systems" approach to cells, by which one can use stimuli to manipulate the output of cellular pathways, promises to revolutionize our understanding and treatment of human disease, provide a robust framework for synthetic biology, and unlock a novel world of nanotechnological possibilities. While phenomenological models for cellular systems can provide great insight into biological pathways, a new level of detail is required to design, perturb, and repair control systems based on cellular architecture. Atomistic simulations are becoming ever more capable of illustrating larger and larger cellular components at meaningful time scales. The utility of such grand simulations for understanding cells, however, is attenuated by a gap in current methodology. Fundamentally, cellular infrastructure operates out of equilibrium, but the theoretical treatment of nonequilibrium systems is far from trivial. As standard simulations are performed under constraints of detailed balance, even a full quantum mechanical treatment of the cell, without modification, would fail to capture the essence of driven biomolecular pathways. How can we connect cellular components to their functional environments and simulate the physics of life?

Weber et al. have recently used space−time perturbation theory (s-ensemble) coupled with MSMs to study nonequilibrium effects in protein folding dynamics.[34,35] In particular, the goal of the study was to identify frozen glassy states in protein dynamics and their role in protein conformational dynamics and function. The effect of the functional environment is taken into account by applying a modifying field or the *s*-field that suppresses or enhances the transitions among states via the following equation $\mathbf{T}(s) = U\,e^{-s} + D$, where T is the titled matrix and U and D are the off-diagonal and diagonal components of the original transitional probability matrix. To study dynamics in the biased s-ensemble, the MSM transition probability matrix is modified to obtain a tilted matrix for a given value of *s*-field and then the eigenvalues of this tilted matrix provide the shifts in the equilibrium probabilities of states in the biased ensemble. Therefore, for high *s*-values ($s > 0$), states with high self-transition probabilities are observed in the conformational ensemble, and for low *s*-values ($s < 0$), faster transitions among the extended conformations are observed. This biased dynamics explains the reasons behind differences in the folding time scales of proteins of similar sequence length. In brief, the folding times for proteins that prefer conformational states with high self-transition probability (glassy dynamics) have slower folding time scales. Surprisingly,

DOI: 10.1021/ar5002999
Acc. Chem. Res. 2015, 48, 414−422

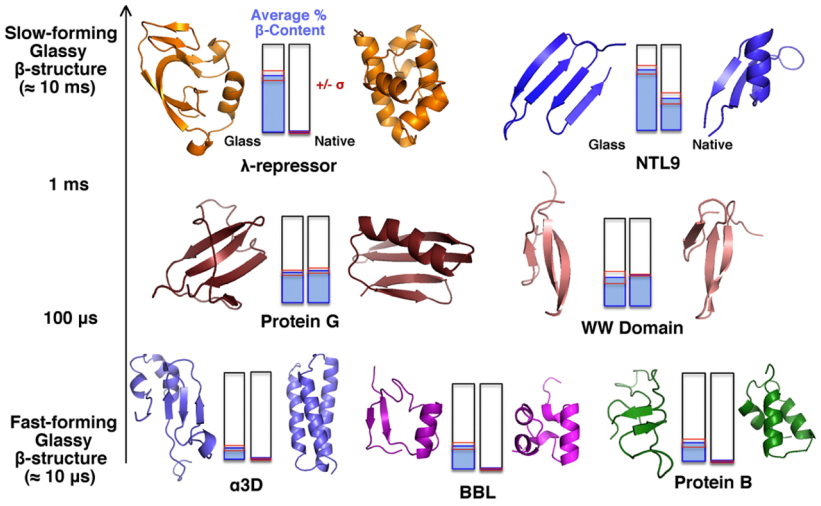


Figure 5. Illustration of protein folding systems' nonnative, amyloid-like glassy states (at left for each system), juxtaposed with their respective native state structures (at right). Time scales related to the formation of glassy β-structure, shown on a log scale (at left), are determined from MFPTs within the respective protein folding MSMs. To suggest the nature of structural fluctuations within the glassy and native states, the colored bars between the structures (on a scale from 0 to 100%) illustrate the mean percentage of β-content in each Markov state, as assigned by the algorithm DSSP, framed by lines that represent ± 1 SD in percentage. Centroids of these states are presented pictorially. Image reproduced with permission from reference 35.

it was found that glassy states tend to have high β-sheet content, indicating a role of glassy dynamics in amyloid fibril formation and stability (Figure 5).

Similarly, specialized statistical mechanical tools offer hope for characterizing dissipative processes in a rigorous fashion. Famous among statistical physicists, relationships like the Crooks fluctuation theorem describe the probabilities of entropy producing (i.e., dissipative) trajectories in general terms. Starting with Lebowitz and Spohn, researchers have beautifully synthesized such fluctuation theorems with the theory of Markov chains so that entropy-producing dynamics can be studied with statistical rigor.[36−38] Recently, we have found that the dominant dissipative trajectories in biased dynamics align well with the *activation pathways* of both GPCRs and kinases. This observation suggests that the molecular machines such as GPCRs and kinases perform meaningful work (signaling) during rare and highly dissipative fluctuations. In other words, the meaningful work involves transitions that dissipate the energy by traversing rare conformational states.

These methodologies could open a new world of possibilities for simulating driven systems. In effect, we have developed an implicit protocol for connecting machinery to its external environment; the possibilities for augmenting simulations of cellular components, fluidic processes, and generic mesoscopic self-assembly with their full, "functional" environments are immense.

Conformational Entropy and Intrinsically Disordered Proteins

Molecular simulations coupled with MSMs have been used to successfully characterize the extensive conformational heterogeneity associated with the dynamics of the intrinsically disordered proteins (IDPs). Recently, MSM-based investigations of IDPs have shed light on the mechanisms of fibril formation for amyloid-β (Aβ), human islet amyloid polypeptide (hIAPP), and other intrinsically disordered peptides.[12,13] Markov state models of the structural ensemble of Aβ_{40} and Aβ_{42} peptides reveal the molecular origins of the higher aggregation propensity of Aβ_{42} compared with that of Aβ_{40}. Lin et al. have also shown how conformational preferences of Aβ change due to the pathogenic mutant E22K (the Italian mutant).[12] Normally, β-sheet formation propensity of the Aβ_{42} peptide changes due to increases in peptide length. The Italian mutation, by contrast, increases the helix formation propensity, which enhances helix−helix interactions between monomers resulting in altered mechanism and kinetics of Aβ aggregation. Similarly, Qiao et al. have found conformations with exposed hydrophobic residues and significant β-sheet content in MSMs of the hIAPP.[13] These conformations could act as a template to induce nucleation of hIAPP fibrils. This mechanism of fibril formation is known as conformational selection, whereby monomer conformations containing pre-existing β-sheet elements selectively collapse and further grow to form fibrils. These observations are consistent not only with several other recent simulations of IDPs but also with experimental results from ion mobility mass spectroscopy. These studies present an ideal example of how MD simulations can provide structural information that is not accessible by experiments and how MSMs can help reduce the complex conformational space exhibited in IDPs into simpler, human-comprehensible models.

To further elaborate upon the theme of IDPs, our group has recently become more focused on order-upon-binding dynamics. Over the course of the past decade, many crystallographic structures have been submitted to the Protein Data Bank containing small IDPs. Several of these structures include a 22 residue fragment of tumor suppressor p53's C-terminal regulatory domain in complex with various binding partners.[39] What is intriguing about these particular structures is that each complex has a unique p53-binding pose and, in some cases, these poses are radically different. Of particular note are S100$\beta\beta$−p53 and sirtuin−p53 complexes, which demonstrate that this same p53 fragment is able to stably bind as both an α-helix and a β-sheet (Figure 6), respectively, and has achieved this in order to inhibit apoptosis in a similar fashion in both.[40,41] How is it that a single peptide is able to promiscuously bind in such a variety of ways? How can elementary models, such as those proposed by Koshland and Fischer, possibly explain these results?[42,43]

DOI: 10.1021/ar5002999
Acc. Chem. Res. 2015, 48, 414−422

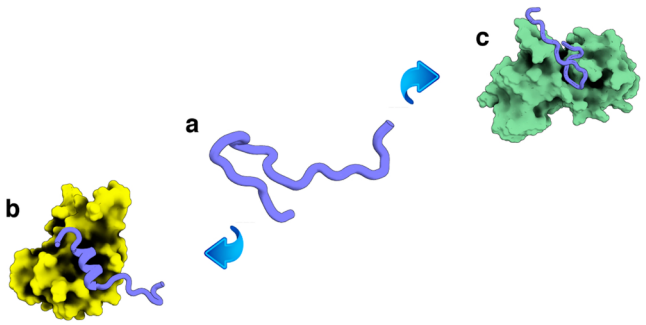


Figure 6. (a) p53 unbound and disordered; (b) p53 bound to S100$\beta\beta$ in a α-helix structure; and (c) p53 bound to sirtuin in a β-sheet structure.

As mentioned previously with Aβ aggregation, conformational selection, in the context of binding, argues that weakly populated transition states are responsible for molecular recognition and subsequent complex formation.[44−46] This is followed by a population shift toward more energetically stable conformations. NMR studies have observed this phenomenon in an IDP fragment of phospholamban binding to protein kinase A.[47] For IDPs, the population shift in conformational selection is thought to be driven by a maximization of the enthalpy of binding due to an increase in the interacting surface area during folding-upon-binding[48] This gain in enthalpy overcomes the conformational entropy loss that is expected as a result of increasing order within the system. However, the specific dynamics and transition states that facilitate these population shifts from unbound-and-disordered to bound-and-ordered remain difficult to determine.

In conjunction with MSM analysis and TPT, atomistic MD has the potential to be an excellent tool to characterize both the kinetics and dynamics of folding-upon-binding for IDPs. Given an appropriate set of order parameters, a series of states of the ligand−target system can be identified from a MSM that represents the binding pathway from unbound-and-unfolded ligand to encounter complexes to bound-and-folded complex, as described in Snow et al.[49] Interestingly, in our studies of p53, MSMs provide evidence for "fly casting", a schema of binding proposed by Shoemaker et al., in the formation of the p53−sirtuin complex but not for binding of p53 to S100$\beta\beta$.[50] Analysis of several high-flux transition pathways reveals that p53 c-terminal regulatory domain forms a transient encounter complex at a distal site on sirtuin before formation of a stably bound β-sheet.[51] By contrast, the same peptide forms the α-helix before binding to S100$\beta\beta$. Both of these results suggest that conformation selection in IDPs, such as p53, can indeed manifest itself through different modes and further highlights the complexity that can be yielded from disorder.

■ LIMITATIONS AND PROSPECTS

In 1990, Karplus and Petsko wrote, "Two limitations in existing simulations are the approximations in the potential energy functions and the lengths of the simulations. The first introduces systematic errors and the second, statistical errors".[52] This observation remains timely because these limitations of MD simulations still represent two of the central challenges in the field. Recent advances in hardware such as graphical processing units[53] (which are now deployed extensively in supercomputer centers), special purpose hardware (Anton),[54] availability of distributed computing platforms (Google Exacycle,[33] Folding@home,[55] Amazon Web Services etc.), and novel sampling algorithms have made the sampling of the long-time-scale phenomena feasible. For example, 500 μs aggregate simulations of kinase catalytic domain reported in a recent study by Shukla et al.[11] could be performed in approximately three months on a cluster with 100 GPUs. Similarly, systematic force-field development procedures and availability of detailed experimental data sets for force field parametrization have yielded better potential energy functions for proteins.[56,57]

However, significant work still needs to be done in order to accurately predict kinetic properties of the protein dynamics and systematic validation of Markov state models. Recently, a first step in this direction has been taken in the form of the framework called "dynamical fingerprints", which has been developed to relate the experimental and MSM-derived kinetic information.[58] Several research groups are now focused on developing protocols to systematically cross-validate the MSM predictions and obtain MSM parameters using an optimization protocol that produces the best estimate of the few slowest dynamics modes of the protein dynamics.[59]

Finally, the exponential growth of sampling ability has led to a deluge of information, which needs to be harnessed into human-comprehendible insights. MSMs provide one way of obtaining such mechanistic insights with broad applications in the field of medicine. Despite its advantages, there are still challenges associated with identifying the best decomposition of conformational space, developing tools for improved error estimation, and creating better approaches to connect MSMs to experimental data. Consequently, this field of inquiry has bright prospects for methodological advances and improving our understanding of the physics of life.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: pande@stanford.edu.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

We thank the NSF (Grant MCB-0954714) and NIH (Grant R01-GM062868) for their support of this work. This work was funded in part by the SIMBIOS NIH National Center for Biomedical Computation through the NIH Roadmap for Medical Research Grant U54 GM07297. D.S. was supported by the Biomedical Data Science Initiative Postdoc Scholar Program of the Stanford School of Medicine. C.X.H. was supported by the NSF Graduate Research Fellowship Program.

Notes

The authors declare no competing financial interest.

Biographies

Diwakar Shukla obtained his Ph.D. in Chemical Engineering from Massachusetts Institute of Technology in 2011. He is currently a postdoctoral fellow in Department of Chemistry at Stanford University. He will join the Department of Chemical Engineering, University of Illinois at Urbana−Champaign, as an assistant professor in January 2015.

Carlos X. Hernández obtained his B.S. in Applied Mathematics from Columbia University. He is currently pursuing a Ph.D. in the

DOI: 10.1021/ar5002999
Acc. Chem. Res. 2015, 48, 414−422

Biophysics Program at Stanford University. His research interests include molecular evolution and intrinsically disordered proteins.

Jeffrey K. Weber obtained his Ph.D. in Chemistry from Stanford University in 2014. He is currently a postdoctoral fellow in the laboratory of Ruhong Zhao at IBM Thomas J. Watson Research Center and Columbia University.

Vijay S. Pande is currently the Director of the Program in Biophysics and a Professor of Chemistry and (by courtesy) of Structural Biology and of Computer Science at Stanford University. He is also the founding Director of the Folding@home Distributed Computing Project.

■ ACKNOWLEDGMENTS

We thank Dr. Morgan Lawrenz, Mohammad M. Sultan, T. J. Lane, Christian R. Schwantes, and Robert S. McGibbon from Department of Chemistry at Stanford University for many insightful discussions about protein dynamics.

■ ABBREVIATIONS

GPCR, G-protein coupled receptor; MSM, Markov state models; HMM, hidden Markov models; MD, molecular dynamics

■ REFERENCES

(1) Abrahamson, E. W.; Ostroy, S. E. The Photochemical and Macromolecular Aspects of Vision. *Prog. Biophys. Mol. Biol.* **1967**, *17*, 179−215.
(2) Lisman, J.; Schulman, H.; Cline, H. The Molecular Basis of CaMKII Function in Synaptic and Behavioural Memory. *Nat. Rev. Neurosci.* **2002**, *3*, 175−190.
(3) Sumner, T. Dazzling History. *Science* **2014**, *343*, 1092−1093.
(4) Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R. G.; Wyckoff, H.; Phillips, D. C. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* **1958**, *181*, 662−666.
(5) Benson, E. E.; Linderstrom-Lang, K. Deuterium Exchange between Myoglobin and Water. *Biochim. Biophys. Acta* **1959**, *32*, 579−581.
(6) Garman, E. F. Developments in X-Ray Crystallographic Structure Determination of Biological Macromolecules. *Science* **2014**, *343*, 1102−1108.
(7) Henzler-Wildman, K.; Kern, D. Dynamic Personalities of Proteins. *Nature* **2007**, *450*, 964−972.
(8) Chodera, J. D.; Noé, F. Markov State Models of Biomolecular Conformational Dynamics. *Curr. Opin. Struct. Biol.* **2014**, *25*, 135−144.
(9) Kohlhoff, K. J.; Shukla, D.; Lawrenz, M.; Bowman, G. R.; Konerding, D. E.; Belov, D.; Altman, R. B.; Pande, V. S. Cloud-Based Simulations on Google Exacycle Reveal Ligand Modulation of GPCR Activation Pathways. *Nat. Chem.* **2014**, *6*, 15−21.
(10) Lane, T. J.; Shukla, D.; Beauchamp, K. A.; Pande, V. S. To Milliseconds and Beyond: Challenges in the Simulation of Protein Folding. *Curr. Opin. Struct. Biol.* **2013**, *23*, 58−65.
(11) Shukla, D.; Meng, Y.; Roux, B.; Pande, V. S. Activation Pathway of Src Kinase Reveals Intermediate States as Targets for Drug Design. *Nat. Commun.* **2014**, *5*, No. 3397.
(12) Lin, Y.-S.; Bowman, G. R.; Beauchamp, K. A.; Pande, V. S. Investigating How Peptide Length and a Pathogenic Mutation Modify the Structural Ensemble of Amyloid Beta Monomer. *Biophys. J.* **2012**, *102*, 315−324.
(13) Qiao, Q.; Bowman, G. R.; Huang, X. Dynamics of an Intrinsically Disordered Protein Reveal Metastable Conformations That Potentially Seed Aggregation. *J. Am. Chem. Soc.* **2013**, *135*, 16092−16101.
(14) Pierce, K. L.; Premont, R. T.; Lefkowitz, R. J. Seven-Transmembrane Receptors. *Nat. Rev. Mol. Cell Biol.* **2002**, *3*, 639−650.
(15) Chodera, J. D.; Pande, V. S. Splitting Probabilities as a Test of Reaction Coordinate Choice in Single-Molecule Experiments. *Phys. Rev. Lett.* **2011**, *107*, No. 098102.
(16) Choudhary, O. P.; Paz, A.; Adelman, J. L.; Colletier, J.-P.; Abramson, J.; Grabe, M. Structure-Guided Simulations Illuminate the Mechanism of ATP Transport through VDAC1. *Nat. Struct. Mol. Biol.* **2014**, *21*, 626−632.
(17) Malmstrom, R. D.; Lee, C. T.; Van Wart, A. T.; Amaro, R. E. Application of Molecular-Dynamics Based Markov State Models to Functional Proteins. *J. Chem. Theory Comput.* **2014**, *10*, 2648−2657.
(18) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything You Wanted to Know about Markov State Models but Were Afraid to Ask. *Methods* **2010**, *52*, 99−105.
(19) Bowman, G. R.; Geissler, P. L. Equilibrium Fluctuations of a Single Folded Protein Reveal a Multitude of Potential Cryptic Allosteric Sites. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 11681−11686.
(20) Prinz, J.-H.; Keller, B.; Noé, F. Probing Molecular Kinetics with Markov Models: Metastable States, Transition Pathways and Spectroscopic Observables. *Phys. Chem. Chem. Phys.* **2011**, *13*, 16912−16927.
(21) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. Progress and Challenges in the Automated Construction of Markov State Models for Full Protein Systems. *J. Chem. Phys.* **2009**, *131*, No. 124101.
(22) E, W.; Vanden-Eijnden, E. Transition-Path Theory and Path-Finding Algorithms for the Study of Rare Events. *Annu. Rev. Phys. Chem.* **2010**, *61*, 391−420.
(23) Bowman, G. R.; Ensign, D. L.; Pande, V. S. Enhanced Modeling via Network Theory: Adaptive Sampling of Markov State Models. *J. Chem. Theory Comput.* **2010**, *6*, 787−794.
(24) Weber, J. K.; Pande, V. S. Characterization and Rapid Sampling of Protein Folding Markov State Model Topologies. *J. Chem. Theory Comput.* **2011**, *7*, 3405−3411.
(25) Zhang, J.; Yang, P. L.; Gray, N. S. Targeting Cancer with Small Molecule Kinase Inhibitors. *Nat. Rev. Cancer* **2009**, *9*, 28−39.
(26) David, C. C.; Jacobs, D. J. Principal component analysis: a method for determining the essential dynamics of proteins. *Methods Mol. Biol.* **2014**, *1084*, 193−226.
(27) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.* **2013**, *9*, 2000−2009.
(28) McGibbon, R. T.; Ramsundar, B.; Sultan, M. M.; Kiss, G.; Pande, V. S. Understanding Protein Dynamics with L1-Regularized Reversible Hidden Markov Models. 2014, arXiv:Q-Bio Stat/1405.1444. arXiv.org e-Print archive. http://arxiv.org/abs/1405.1444.
(29) Perez-Hernandez, G.; Paul, F.; Giorgino, T.; de Fabritiis, G.; Noé, F. Identification of Slow Molecular Order Parameters for Markov Model Construction.2013 arXiv:Phys. Q-Bio/1302.6614. arXiv.org e-Print archive. http://arxiv.org/abs/1302.6614.
(30) Razavi, A. M.; Wuest, W. M.; Voelz, V. A. Computational Screening and Selection of Cyclic Peptide Hairpin Mimetics by Molecular Simulation and Kinetic Network Models. *J. Chem. Inf. Model.* **2014**, *54*, 1425−1432.
(31) Deupi, X.; Standfuss, J. Structural Insights into Agonist-Induced Activation of G-Protein-Coupled Receptors. *Curr. Opin. Struct. Biol.* **2011**, *21*, 541−551.
(32) Trzaskowski, B.; Latek, D.; Yuan, S.; Ghoshdastider, U.; Debinski, A.; Filipek, S. Action of Molecular Switches in GPCRs−Theoretical and Experimental Studies. *Curr. Med. Chem.* **2012**, *19*, 1090−1109.
(33) Hellerstein, J. L.; Kohlhoff, K. J.; Konerding, D. E. Science in the Cloud: Accelerating Discovery in the 21st Century. *IEEE Internet Comput.* **2012**, 64−68.
(34) Weber, J. K.; Jack, R. L.; Pande, V. S. Emergence of Glass-like Behavior in Markov State Models of Protein Folding Dynamics. *J. Am. Chem. Soc.* **2013**, *135*, 5501−5504.
(35) Weber, J. K.; Jack, R. L.; Schwantes, C. R.; Pande, V. S. Dynamical Phase Transitions Reveal Amyloid-like States on Protein Folding Landscape. *Biophys. J.* **2014**, *107* (4), 974−982.

DOI: 10.1021/ar5002999
Acc. Chem. Res. 2015, 48, 414−422

(36) Lebowitz, J. L.; Spohn, H. A Gallavotti−Cohen-Type Symmetry in the Large Deviation Functional for Stochastic Dynamics. *J. Stat. Phys.* **1999**, *95*, 333−365.
(37) Crooks, G. E. Path-Ensemble Averages in Systems Driven far from Equilibrium. *Phys. Rev. E* **2000**, *61*, 2361−2366.
(38) Jarzynski, C. Nonequilibrium Equality for Free Energy Differences. *Phys. Rev. Lett.* **1997**, *78*, 2690−2693.
(39) Uversky, V. N.; Oldfield, C. J.; Dunker, A. K. Intrinsically Disordered Proteins in Human Diseases: Introducing the D^2 Concept. *Annu. Rev. Biophys.* **2008**, *37*, 215−246.
(40) Avalos, J. L.; Celic, I.; Muhammad, S.; Cosgrove, M. S.; Boeke, J. D.; Wolberger, C. Structure of a Sir2 Enzyme Bound to an Acetylated p53 Peptide. *Mol. Cell* **2002**, *10*, 523−535.
(41) Rustandi, R. R.; Baldisseri, D. M.; Weber, D. J. Structure of the Negative Regulatory Domain of p53 Bound to S100B($\beta\beta$). *Nat. Struct. Biol.* **2000**, *7*, 570−574.
(42) Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **1958**, *44*, 98−104.
(43) Fischer, E. Einfluss der Configuration auf die Wirkung der Enzyme. *Ber. Dtsch. Chem. Ges.* **1894**, *27*, 2985−2993.
(44) Boehr, D. D.; Nussinov, R.; Wright, P. E. The Role of Dynamic Conformational Ensembles in Biomolecular Recognition. *Nat. Chem. Biol.* **2009**, *5*, 789−796.
(45) Gibbs, A. C. Elements and Modulation of Functional Dynamics. *J. Med. Chem.* **2014**, *57*, 7819−7837.
(46) Straub, F. B.; Szabolcsi, G. O Dinamicseszkij Aszpektah Sztukturü Fermentov (On the Dynamic Aspects of Protein Structure). *Molecular Biology, Problems and Perspectives*; Braunstein, A. E., Ed.; Izdat. Nauka: Moscow, 1964.
(47) Masterson, L. R.; Cheng, C.; Yu, T.; Tonelli, M.; Kornev, A.; Taylor, S. S.; Veglia, G. Dynamics Connect Substrate Recognition to Catalysis in Protein Kinase A. *Nat. Chem. Biol.* **2010**, *6*, 821−828.
(48) Flock, T.; Weatheritt, R. J.; Latysheva, N. S.; Babu, M. M. Controlling Entropy to Tune the Functions of Intrinsically Disordered Regions. *Curr. Opin. Struct. Biol.* **2014**, *26*, 62−72.
(49) Snow, C. D.; Rhee, Y. M.; Pande, V. S. Kinetic Definition of Protein Folding Transition State Ensembles and Reaction Coordinates. *Biophys. J.* **2006**, *91*, 14−24.
(50) Shoemaker, B. A.; Portman, J. J.; Wolynes, P. G. Speeding Molecular Recognition by Using the Folding Funnel: The Fly-Casting Mechanism. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 8868−8873.
(51) Ubbink, M. The Courtship of Proteins: Understanding the Encounter Complex. *FEBS Lett.* **2009**, *583*, 1060−1066.
(52) Karplus, M.; Petsko, G. A. Molecular Dynamics Simulations in Biology. *Nature* **1990**, *347*, 631−639.
(53) Eastman, P.; Friedrichs, M. S.; Chodera, J. D.; Radmer, R. J.; Bruns, C. M.; Ku, J. P.; Beauchamp, K. A.; Lane, T. J.; Wang, L.-P.; Shukla, D.; Tye, T.; Houston, M.; Stich, T.; Klein, C.; Shirts, M. R.; Pande, V. S. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J. Chem. Theory Comput.* **2013**, *9*, 461−469.
(54) Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Eastwood, M. P.; Ierardi, D. J.; Klepeis, J. L.; Kuskin, J. S.; Larson, R. H.; Lindorff-Larsen, K.; Maragakis, P.; Moraes, M. A.; Piana, S.; Shan, Y.; Towles, B. Anton, a Special-Purpose Machine for Molecular Dynamics Simulation. In *Proceedings of the 34th Annual International Symposium on Computer Architecture, ISCA '07*; ACM: New York, 2007; pp 1−12.
(55) Shirts, M.; Pande, V. S. Screen Savers of the World Unite! *Science* **2000**, *290*, 1903−1904.
(56) Paulechka, E.; Kroenlein, K.; Kazakov, A.; Frenkel, M. A Systematic Approach for Development of an OPLS-Like Force Field and Its Application to Hydrofluorocarbons. *J. Phys. Chem. B* **2012**, *116*, 14389−14397.
(57) Wang, L.-P.; Chen, J.; Van Voorhis, T. Systematic Parametrization of Polarizable Force Fields from Quantum Chemistry Data. *J. Chem. Theory Comput.* **2013**, *9*, 452−460.
(58) Noé, F.; Doose, S.; Daidone, I.; Löllmann, M.; Sauer, M.; Chodera, J. D.; Smith, J. C. Dynamical Fingerprints for Probing Individual Relaxation Processes in Biomolecular Dynamics with Simulations and Kinetic Experiments. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 4822−4827.
(59) McGibbon, R. T.; Pande, V. S. Variational Cross-Validation of Slow Dynamical Modes in Molecular Kinetics. 2014, arXiv:Phys. Q-Bio Stat/1407.8083. arXiv.org e-Print archive. http://arxiv.org/abs/1407.8083.

DOI: 10.1021/ar5002999
Acc. Chem. Res. 2015, 48, 414−422