JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Article

pubs.acs.org/JACS

Timing Correlations in Proteins Predict Functional Modules and Dynamic Allostery

Milo M. Lin*

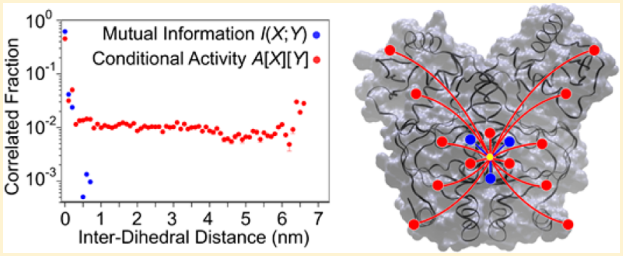
Green Center for Molecular, Computational, and Systems Biology, University of Texas Southwestern Medical Center, Dallas, Texas 75390, United States

Department of Biophyics, University of Texas Southwestern Medical Center, Dallas, Texas 75390, United States

Pitzer Center for Theoretical Chemistry, University of California, Berkeley, California 94720, United States

Supporting Information

ABSTRACT: How protein structure encodes functionality is not fully understood. For example, long-range intraprotein communication can occur without measurable conformational change and is often not captured by existing structural correlation functions. It is shown here that important functional information is encoded in the timing of protein motions, rather than motion itself. I introduce the conditional activity function to quantify such timing correlations among the degrees of freedom within proteins. For three proteins, the conditional activities between side-chain dihedral angles were computed using the output of microseconds-long atomistic simulations. The new approach demonstrates that a sparse fraction of side-chain pairs are dynamically correlated over long distances (spanning protein lengths up to 7 nm), in sharp contrast to structural correlations, which are short-ranged (<1 nm). Regions of high self- and inter-side-chain dynamical correlations are found, corresponding to experimentally determined functional modules and allosteric connections, respectively.



INTRODUCTION

Proteins perform most of the functional and infrastructural roles of life at the subcellular level. A protein molecule is a nanometer-scale polymer of amino acids, which typically folds into a unique native conformation encoding a specific functional role.[1] The functionality in one region of a protein (or protein complex) can sometimes be switched on/off or modulated by perturbing (e.g., via ligand binding) a distant part of the protein; this property is known as allostery.[2,3] Intriguingly, there is growing evidence for allostery occurring without discernible change in the dominant protein conformation(s) following the perturbation. This more subtle means of communication, termed dynamical allostery,[4] shows that proteins can communicate by merely altering the spread and time scales of fluctuations around and between the dominant conformations.[5] Recently, the focus has shifted from identifying specific structural transitions to characterizing shifts in the conformational ensemble,[6,7] with explicit-atom simulations playing a key role in providing the necessary spatial and temporal resolution to describe the subtle conformational fluctuations involved.[8,9] Because the dynamics of a protein's conformational probability distribution determines its function, information about a protein's functionality should in principle be encoded by its structure. Yet, the link between structure and function remains an area of active research, especially in regards to elucidating intraprotein communication pathways.

Binding of the allosteric ligand can distort the probability distribution of a protein in conformational space or time without changing the most probable conformation; this can occur via changing the width of the probability distributions of conformations (i.e., the entropy) as well as interconversion frequencies (i.e., waiting times between conformational changes).[10] In contrast to existing approaches, which deal with the former by characterizing correlated structural changes, this work shows that a crucial missing piece of the connection between structure and function lies in the latter: namely, correlated changes in the waiting times.

To illustrate the role of timing, consider the native conformation of a protein, a tightly packed amorphous (i.e., nonperiodic) material in which motion of one part of the protein often requires concerted motions of other parts of the protein. If such a system is sufficiently densely packed, an immobilization transition can occur in which some parts of the system become jammed despite isolated regions of dynamical mobility.[11] There are usually no structural signatures that can identify what parts of the system are jammed. In such systems, what matters for characterizing the system is not the direction that a particle can move but whether and when a particle can move. Indeed, the correlation of a particle's current level of dynamical activity with its dynamical history (i.e., non-Poissonian dynamics) is a signature of immobilization behavior. Such a signature has long been associated with glass-forming

Received: August 7, 2015
Revised: August 19, 2015
Published: March 22, 2016

DOI: 10.1021/jacs.5b08814
J. Am. Chem. Soc. 2016, 138, 5036−5043

liquids,[12] and, remarkably, was recently found to describe side chain dynamics within proteins.[13] These results suggest the possibility that, by selectively colocalizing the jammed regions, proteins can contain intramolecular pathways that are synchronized in their fluctuations. This motivates the formulation of a new type of function that measures correlations in timing. In this paper, I define such a function, which I call the conditional activity, and use it to map regions of dynamical memory and predict long-range dynamical correlations in proteins. It is shown that such information is not revealed by thermodynamic correlation functions like the mutual information.

The Conditional Activity Function. The standard approach to elucidate the collective behavior of complex and structurally heterogeneous systems such as proteins is to calculate correlations between different parts of the system. The most widely used correlation function is the mutual information,[14] which for observables X and Y is the entropy of X plus the entropy of Y minus the entropy of the total system comprised of X and Y: $I(X;Y) = \Sigma_{x\in X}P(x)\ \ln[P(x)] + \Sigma_{y\in Y}P(y)\ \ln[P(y)] - \Sigma_{x\in X}\Sigma_{y\in Y}P(x,y)\ \ln[P(x,\ y)]$, where P is the probability function. Intuitively, this is the amount of entropy (or information) that X shares with Y; $I(X;Y)$ is therefore also the reduction in entropy of X if the state of Y is known. Throughout, I use the natural logarithm in defining entropies; translation to the conventions of information theory and thermodynamics involve trivial unit conversions via multiplication by ln2 and Boltzmann's constant, respectively.

The entropy can be defined as the logarithm of the number of effective or likely states (which is equal to the total number of states if all accessible states have equal probability, such as in the microcanonical ensemble). Let $\Omega(X)$ and $\Omega(X|Y)$ represent the number of effective states of X and the number of states of X given that Y is known, respectively. The mutual information can be expressed as $\ln[\Omega(X)] - \ln[\Omega(X|Y)]$, which can be rewritten as a logarithm of the ratio between the conditional and unconditional state space sizes:

$$I(X;\ Y) = -\ln\left[\frac{\Omega(X|Y)}{\Omega(X)}\right] \tag{1}$$

The mutual information is a thermodynamic function and, for an ergodic system, can be obtained by averaging over an ensemble of independent instances of the system rather than averaging a single system over time. Below, I introduce the conditional activity function to measure correlations in timing, and for which time-averaging is not equivalent to ensemble-averaging.

The conditional activity function, formulated here, is well-defined for systems in which the degrees of freedom can be characterized (or well-approximated) as residing in one of a finite set of discrete states. If this is the case, the conditional activity function between any pair of degrees of freedom is directly calculated from the time series of the degrees of freedom as follows. Let X and Y denote two degrees of freedom that can transition between distinct states over time. Define the transition time function, $T(X,\ i)$, to be the time of the ith transition of X. Define the waiting time function, $W(X,\ t)$, to be the time interval, starting at time t, until the next transition of X. These two functions do not require the status of each degree of freedom at each time point; they simply require a list of times at which each degree of freedom makes a transition. Let the number of total transitions of X be $N(X) \gg 1$ during the observation time $\tau \equiv T(X,\ N(X))$. Then the mean persistence time of X is

$$\tau_p[X] = \frac{1}{2\tau}\sum_{i=1}^{N(X)} W(X,\ T(X,\ i))^2 \tag{2}$$

which is proportional to the mean-squared waiting times between transitions in the time series of X. The squared waiting times are the product of each waiting time multiplied by the probability of selecting that waiting time, which is itself proportional to that waiting time; intuitively, the mean persistence time is the expected waiting time, starting from a random time point, until the next transition of X.

The exchange time of X following the $(i + 1)$th transition in Y is $W(X,\ T(Y,\ i + 1))$. The probability that, starting from a randomly chosen initial observation time, the $(i + 1)$th transition of Y is first observed is given by $\frac{1}{\tau}W(Y,\ T(Y,\ i))$, which is simply the time interval from the ith to the $(i + 1)$th transition of Y, divided by the total trajectory time. Therefore, the mean exchange time of X following a transition in Y is

$$\tau_x[X][Y] = \frac{1}{\tau}\sum_{i=1}^{N(Y)-1} W(X,\ T(Y,\ i + 1))W(Y,\ T(Y,\ i)) \tag{3}$$

Intuitively, this corresponds to observing the process starting at a random time point, waiting for the first instance of a transition in Y, measuring the subsequent time it takes for a transition to occur in X, and averaging this measured time over many repeats of this process.

Each term in the summations of eqs 2−3 is a waiting time of X multiplied by the (un-normalized) probability of encountering that waiting time following either a previous transition of X (eq 2) or a transition in Y (eq 3); the reciprocal of the observation time in front of the summation normalizes the waiting time probability. For $X = Y$, these relations correspond to the mean of the persistence and exchange time functions encountered in the theory of glasses.[11,15] These are both measures of the waiting time until the next event, and differ in their definition of starting times: the persistence time uses a random time-point whereas the exchange time uses a time-point at which a previous event occurred. Moreover, previous work[13] on protein side chain dynamics has shown that these time scales can differ dramatically, demonstrating intermittency and correlation of protein side-chain displacement. The exchange time is generalized in this work to include the waiting time of one degree of freedom conditional on (i.e., after) the change in a different degree of freedom. The conditional activity of X on Y is defined to be

$$A[X][Y] \equiv -\ln\left[\frac{\tau_x[X][Y]}{\tau_p[X]}\right] \tag{4}$$

which is analogous to the expression for mutual information (eq 1), except that the conditioned and unconditioned *number* of states in eq 1 are replaced by the exchange and persistence *times*, respectively, in eq 4. If X and Y are independent processes, then $A[X][Y] = 0$. For the special case of $X = Y$, $\tau_x[X][X]$ is the (normalized) mean product of adjacent waiting times in the time series of X. The diagonal elements $A[X][X]$ represent the dynamical memory of X; if the waiting time until the next transition of X is independent of the previous history of X, $A[X][X] = 0$, meaning that the transition statistics of X are

DOI: 10.1021/jacs.5b08814
J. Am. Chem. Soc. 2016, 138, 5036−5043

Poissonian (i.e., "memoryless"). Note that, unlike the mutual information, the conditional activity matrix is generally not symmetric: $A[X][Y]$ is in general not equal to $A[Y][X]$. This is because, unlike the "and" operator used to construct the mutual information, the "after" operator used to construct the conditional activity is not commutative: time is directional.

RESULTS

The degrees of freedom studied here are side-chain dihedral angles. Side-chain fluctuations have been shown to play a critical role in protein function,[16] and recent experimental[17,18] and computational[19,20] work demonstrated that such fluctuations occur throughout the protein structure, even in the densely packed protein core. Such interconversions can be correlated between different regions of proteins despite minimal backbone mobility.[21,22] Furthermore, side-chain dynamical statistics can provide a natural reporter for the parts of the protein that are jammed, even if the underlying mechanism involves subtle backbone conformational changes. Figure 1 shows the conformational statistics of a few

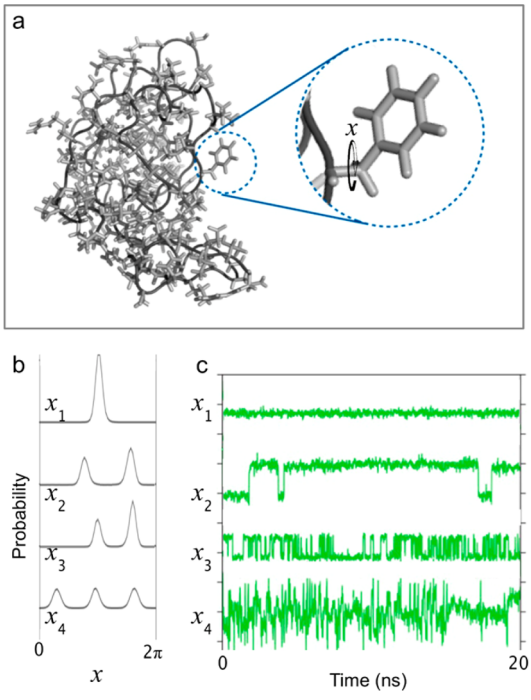


Figure 1. Protein side-chain dynamics. The degrees of freedom correspond to rotations of residue dihedral angles denoted by χ (a). Despite having similar probability distributions (b), degrees of freedom may undergo very different temporal dynamics (c).

representative side-chain dihedral angles in a typical globular protein, showing that the dihedral angle distribution is well-represented by a set of distinct basins. Despite similar (thermodynamic) occupancies showing typically three or two minimal-energy dihedral angle basins (Figure 1b), the (dynamical) waiting time distribution can vary over many orders of magnitude for a given dihedral angle, as well as among different dihedrals (Figure 1c).

I calculated the conditional activity of the side-chain dynamics of catabolite activator protein, which is a homodimeric protein complex. Catabolite activator protein binds to the small signaling molecule cAMP, which is produced in response to low glucose levels; upon binding cAMP, catabolite activator protein binds to DNA and recruits RNA polymerase to transcribe genes for metabolism of energy sources other than glucose. There is dynamical allostery between two cAMP-binding pockets and the DNA-binding sites of the protein complex.[4]

Time-series of side-chain dihedral angle transitions were obtained from all-atom molecular dynamics simulations of catabolite activator protein starting from the X-ray crystallographic structure.[23] Crucially, DNA and cAMP were absent from the simulations in order to calculate the conformational dynamics and network behavior intrinsic to the protein complex. The protein complex was equilibrated by coupling to a 28 °C temperature and 1 atm pressure bath, in a single simulation trajectory lasting three microseconds. In addition to side-chain dihedral angles, the backbone conformation was also monitored for structural signatures such as root-mean-square-deviation (RMSD) analysis, both on a per-residue level, as well as for the protein complex as a whole to monitor convergence (Figure S1). Normal modes of the complex were also calculated (Movies S1–S4). These analyses showed that the loop region of the complex near the active sight is highly flexible and elucidated large-scale conformational movements, but did not reveal the residue-specific correlations obtained from analyzing side-chain dihedral statistics.

From the simulation trajectory, the conditional activity was calculated according to eq 4. For all side-chains studied, the dihedral angles fall into one, two or three distinct rotameric basins, each centered at a peak in the dihedral population distribution (see Figure 1). For each of the 850 distinct side-chain dihedral angles, the list of times when the dihedral transitioned between basins was calculated using the simulation trajectory. Then, the persistence and exchange times were calculated as in eqs 2–3, by applying the transition time and waiting time functions on the list of times. The mutual information between all dihedral angles was also calculated in order to compare and contrast with the conditional activity function.

The diagonal elements of the mutual information matrix are the conformational entropies of each side chain dihedral angle; the exponential of the entropy corresponds to the effective number of rotameric states available to the dihedral angle. The side-chain conformational entropy is uniform over the surface of the protein, and suppressed in the protein interior due to the predominance of residues occupying a single dihedral basin in the tightly packed interior (see Figure 2a).

In an analogous fashion, the diagonal elements of the conditional activity matrix measure the dynamical memory (i.e., the extent of non-Poissonian dynamics) of each side chain dihedral angle; if the dynamics can be well-characterized by a fast rate in an activated state and a slow rate of activation (unjamming), the exponential of the dynamical memory corresponds to the ratio of these two rates. Figure 2b shows the dihedral angles as spheres whose volumes are proportional to the dynamical memory. Whereas some degrees of freedom have Poissonian waiting-time statistics (negligible dynamical memory), for others the expected waiting time until the next transition is reduced by more than 2 orders of magnitude following a previous transition; in the latter case, the dynamical transitions are bunched into short periods of high activity followed by relatively long quiescent time periods. This heterogeneous spatial distribution of dynamical memory is in contrast to the homogeneous entropy distribution (Figure 2a). It is important to note that the dynamical memory of dihedrals

DOI: 10.1021/jacs.5b08814
J. Am. Chem. Soc. 2016, 138, 5036–5043

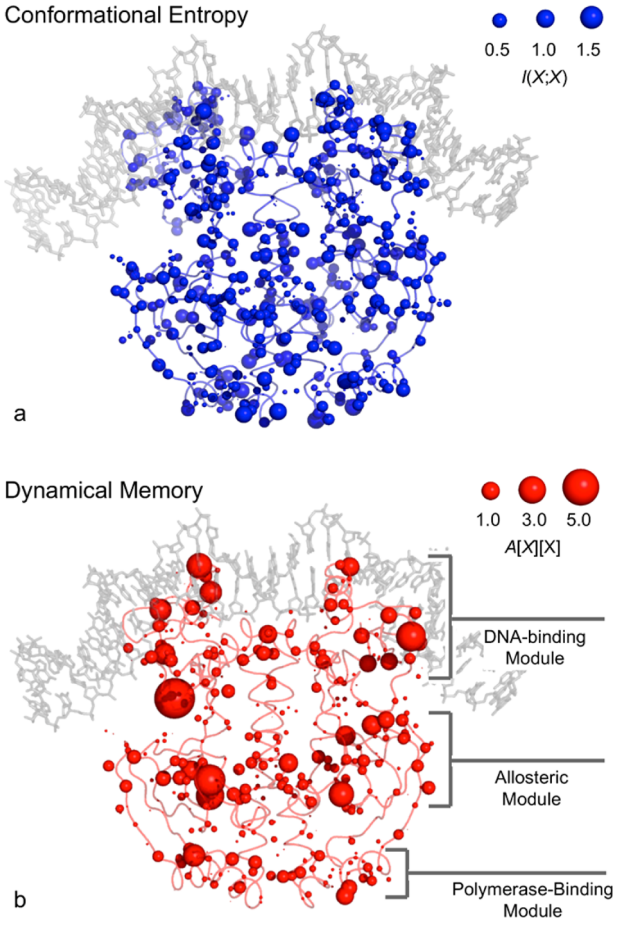


Figure 2. Conformational entropy versus dynamical memory in catabolite activator protein. The diagonal of the mutual information matrix $I(X;X)$ is the conformational entropy of dihedral angle X (blue; a). The diagonal of the conditional activity matrix $A[X][X]$ is the dynamical memory of dihedral angle X (red, b). Each degree of freedom is drawn as a sphere with the sphere volume proportional to the correlation function. The entropy is homogeneous in both magnitude and spatial distribution (a), whereas the dynamical memory is heterogeneous in both aspects. Three distinct regions with high dynamical memory correspond to the three functional modules: DNA-binding module, allosteric cAMP binding module, and RNA polymerase binding module (b). The dihedral angles with the most dynamical memory undergo dynamics in bursts, with intraburst waiting times hundreds of times shorter than interburst waiting times.

is due to the collective interactions with the other degrees of freedom in the protein; however, the dynamics of the protein in the full (high-dimensional) microstate space is always memoryless; the memory effect arises from projecting this full space onto a lower-dimensional space such as that of an individual dihedral.

The dynamical memory is clustered into three separate modules corresponding to the top (DNA binding), middle (allosteric binding), and bottom (RNA polymerase binding) parts of the protein.[24] This contrasts with the uniform distribution of conformational entropy. These contiguous regions of high dynamical memory correspond to the experimentally determined functional modules, yet possess no structural signatures to differentiate them from the nonfunctional regions of the protein. The dihedrals with high levels of dynamical memory are clustered together in functional modules because those are the regions of the protein whose side-chain fluctuations are synchronous with that of their neighbors in the module. However, different subsets of the module may be dynamically correlated at any one time: the dynamical memory is not interaction-specific. In contrast, specific correlations among the dihedrals, which may correspond to robust pathways of intraprotein communication, are contained in the off-diagonal elements of the conditional activity matrix (see below).

Dynamical correlations between different degrees of freedom are encoded by the off-diagonal elements of the conditional activity matrix. Shown in Figure 3a is a comparison of how the interdihedral mutual information and conditional activity decay as a function of interdihedral distance r. Given a correlation function, $f(r)$, the fraction of correlated dihedral pairs is the fraction of dihedral pairs separated by r for which $f(r)/f(0) > 1/e \approx 0.368$ (e-fold decay). The correlation functions of interest are the conditional activity and mutual information for any pair

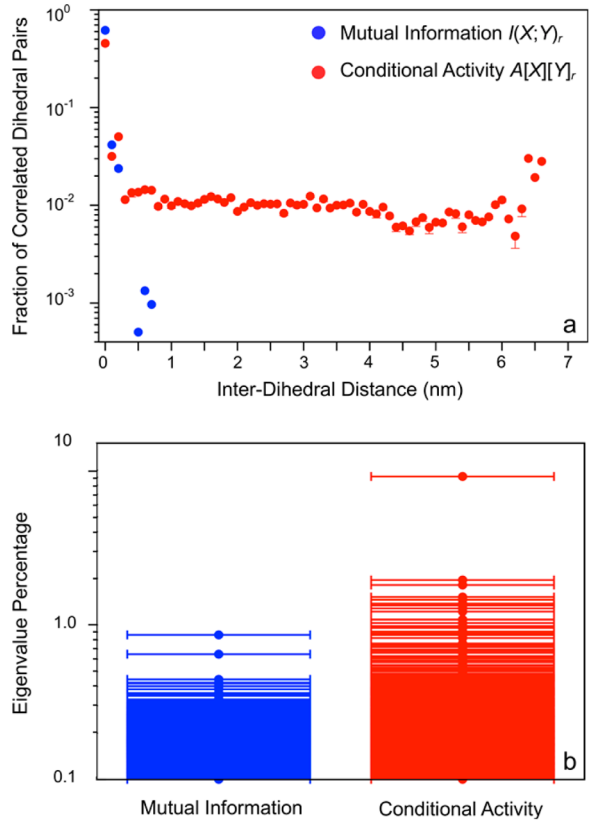


Figure 3. Interdihedral mutual information vs conditional activity. Fraction of correlated dihedral pairs as a function of distance separation in catabolite activator protein (a). The mutual information (thermodynamic correlations; blue) between dihedral angles is negligible beyond 0.75 nm. In contrast, at least 1% of pairwise conditional activities (dynamic correlations; red) persist, irrespective of separation distance. The error bars were obtained as described in the text. Eigenvalue spectrum of pairwise dihedral correlations in catabolite activator protein (b). The entropy is evenly spread among the eigenvalues of the mutual information (blue), with no eigenvalue accounting for more than 1% of the total dihedral entropy. In contrast, the principal eigenvalue of the conditional activity accounts for 10% of all dynamical memory in the protein complex, and is well separated from the other eigenvalues. The eigenvectors corresponding to the top two eigenvalues of the mutual information, as well as the principal eigenvalue of the conditional activity, are shown in Figures S2 and 4c, respectively.

DOI: 10.1021/jacs.5b08814
J. Am. Chem. Soc. 2016, 138, 5036−5043

of degrees of freedom separated by distance r: $f(r) \in \{A[X][Y]_r, I(X;Y)_r\}$. The error bars are set by the number of sampling events in the simulation time window (see Methods section). For most pairs of correlated dihedrals, the correlation signal is one to two orders of magnitude larger than the expected noise. The mutual information decays on the length scale of 0.5 nm and no pair of dihedrals separated by more than a nanometer share any measurable mutual entropy. In contrast, for all interdihedral distances greater than 0.4 nm, 3011 dihedral pairs are dynamically correlated, corresponding to about 1% of all dihedral pairs. Remarkably, over up to 6.5 nm, there is no discernible dampening of fraction of dihedral pairs that are dynamically correlated; at physiological levels of thermal noise, dynamical correlations seem to be capable of being transmitted over much longer distances than thermodynamic correlations. This finding is consistent with the picture that a subset of backbone conformational modes, which can include long-ranged collective motion, control the dynamical activity of multiple dihedral motions, leading to correlated periods of dynamical motion even though the motions themselves are not structurally correlated (i.e., coherent).

The interdihedral timing correlations are not only longer-ranged, but also can be concentrated into strongly correlated modes. The mutual information matrix and symmetrized conditional activity matrix (mean of the matrix with its transpose) were diagonalized, as done in earlier studies,[21,25] to elucidate to what extent the correlations are concentrated in a few principal communities of dihedrals and the spatial extent of such communities. Figure 3b shows the eigenvalue spectra of the mutual information and conditional activity matrices. The principal eigenvalue comprises 10% of all positive timing correlations in the protein complex despite being one of 850 total modes, and is qualitatively separated from the rest of the eigenvalues by a gap in the eigenvalue spectrum. In constrast, there are two principal eigenvalues of the mutual information matrix, neither of which is significantly separated from the rest of the spectrum, and no mode accounts for more than 1% of the total mutual information.

Figures 2 and 3 highlight two features of timing correlations that are not captured by structural correlations. First, from Figure 2, comparing the diagonal of the conditional activity matrix with that of the mutual information, we observe that there can be spatial heterogeneity of the former (non-Poissonian behavior) despite a homogeneous distribution of the latter (entropy); a consequence of this feature is that this heterogeneity can be used to locate regions of the protein whose dihedrals are highly sensitive to their environment and are thus good candidates for being functional modules, as is the case for catabolite activator protein. Second, the off-diagonal (interdihedral) elements of the conditional activity are correlated over much larger distances than the mutual information (Figure 3a). In addition, certain communities of dihedrals have a disproportionate share of the dynamical correlation (Figure 3b). A consequence of these features is the possibility to encode dynamical allostery, as demonstrated below.

Figure 4c shows the principal eigenvector of the symmetrized conditional activity matrix, whose eigenvalue is well separated from all others as shown in Figure 3b, mapped onto the structure of catabolite activator protein. The principal eigenvector of the conditional activity identifies the set of dihedrals that are maximally intercorrelated in terms of being active versus inactive. The sphere size of each dihedral angle is

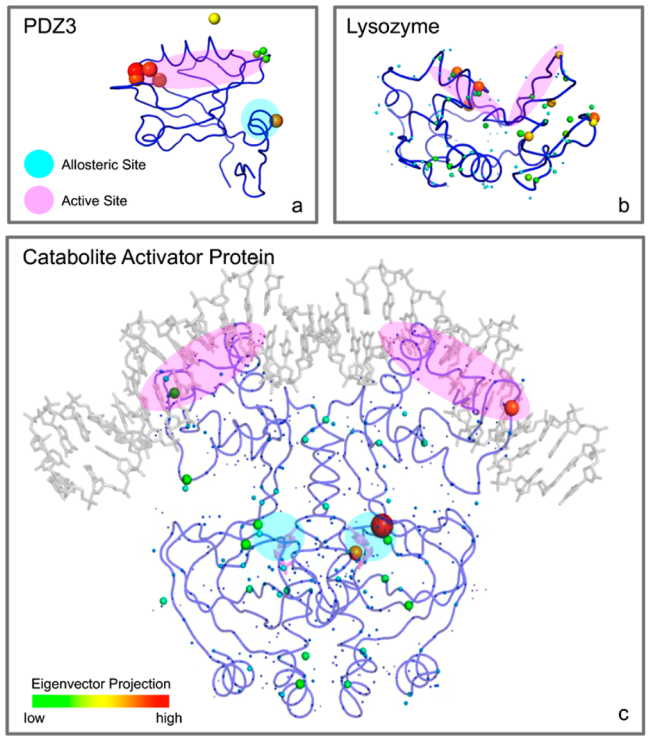


Figure 4. Dynamically correlated subnetworks. In the allosteric proteins (a and c), the principal eigenvector of the (symmetrized) conditional activity matrix predicts a sparse subset of degrees of freedom that are dynamically connected (sphere size proportional to degree of representation in the principal eigenvector); this subset pinpoints the experimentally determined allosteric site (cyan highlight) and includes the active site (magenta highlight). In contrast, for the nonallosteric lysozyme (b), the principal eigenvector is spread over the entire protein.

proportional to the contribution of the dihedral to the principal eigenvector. The previously known (experimentally determined) active and allosteric sites are shaded in magenta and cyan, respectively. The simulations started from the X-ray structure reveal that dynamical correlations are encoded between the cAMP binding site of one monomer of the homodimer complex and the active site of the other protein monomer. Because the two monomers are identical, on time scales longer than the three-microsecond simulation window, the protein dimer is expected to interconvert to an alternate conformational basin in which the other allosteric-active site pair becomes correlated. This structure–function relationship is not revealed in the two principal eigenvectors of the mutual information, which are localized and do not overlap with dihedrals in the active site (Figure S2). Their corresponding eigenvalues are not well separated from the rest of the spectrum (Figure 3b) and so the communities of dihedrals they represent are not significantly more thermodynamically coupled than other modes of the mutual information matrix.

To further validate the ability of the conditional activity to identify selective coupling interactions and the correspondence to dynamical allostery, two other proteins were also simulated in the same manner for two microseconds each: PDZ3[26] and human lysozyme, with 228 and 246 side chain dihedral angles, respectively.[27] PDZ3 (PSD-95/SAP90) is a binding domain that, unlike other PDZ domains, possesses an additional tail helix; there is dynamical allostery between the tail helix and the binding groove;[28] these regions are correctly captured by the

DOI: 10.1021/jacs.5b08814
J. Am. Chem. Soc. 2016, 138, 5036–5043

principal eigenvector of the conditional activity matrix (Figure 4a). Lysozyme, which served as a control, does not possess long-range allostery; in correspondence, the principal eigenvector is evenly distributed throughout the protein (Figure 4b). For all three proteins studied, the principal eigenvalue contains between 10 and 30% of the total dynamical memory. This indicates that the existence and long-range correlations of waiting times between side chain fluctuations is intrinsic to proteins. However, only the two dynamically allosteric proteins (catabolite activator protein and PDZ3) are folded in such a way as to concentrate these correlations into sparse regions that correspond to regions of allosteric control.

DISCUSSION

It has long been understood that allostery can occur without changing protein conformation.[7,9] Whereas attention has up to now focused on explaining allostery using equilibrium thermodynamic quantities such as enthalpy and entropy (including mutual entropy), these results show that non-equilibrium information (namely, correlations between side-chain dihedral rotation waiting times) is quantified by a new measure: the conditional activity. The conditional activity can reveal the distinct functional modules within proteins as well as the sparse networks of correlated dynamics that can enable allostery without conformational change. Like any other correlation function, the conditional activity does not establish causality. However, it shows that long-distance timing correlations can exist even when enthalpic and entropic interactions are short-ranged.

As shown by Figure 3a, the inter-residue mutual information attenuates within one nanometer, indicating that structural correlations do not extend beyond nearest neighbor. From the connectivity analysis[29,30] of the mutual information network, most dihedrals that share mutual information do so with more than one neighbor. Thus, the mutual information is short-ranged not because the nearest-neighbor structural correlations are too sparse to form pathways. Rather, it is because the structural correlations lose coherence beyond nearest-neighbor due to thermal noise: direct structural correlations between side-chains are not transitive. However, it is important to note that long-range side-chain structural correlations can be mediated by backbone conformational changes, for example in nondynamical allostery. Therefore, the timing-based approach of the conditional activity complements, rather than replaces, the existing structure-based methods.

The possibility remains that the backbone is involved in transmitting the dynamical correlations; if so, such rearrangements are of the same small magnitude as other backbone fluctuations that do not give rise to enhanced/suppressed side-chain activity, and are therefore not detectable by directly monitoring backbone motions. For example, neither the backbone RMSD fluctuations (Figure S1), nor elastic network analysis of the protein complex (Movies S1−S4) reveal either the three-module functional compartmentalization or the allosteric connection between the active site and ligand-binding pocket.

These findings complement the large body of experimental work on allostery without backbone conformational change. For catabolite activator protein, entropic effects were found to govern cooperative binding between the first and second cAMP ligands, with amide exchange rates showing that the second ligand binding event significantly increased the rigidity of the central helix.[4] In terms of the allostery between cAMP and DNA binding, mutating various key residues in catabolite activator protein, for example at residues 136 and 138, has been experimentally implicated in modulating the allosteric effect of cyclic AMP on DNA binding.[31] These residues are at or near the hinge connecting the central helices to the DNA binding module; as such, mutating them could have important consequences for the rigidity of the hinge and therefore the dynamical coupling between the allosteric region and the DNA binding region, consistent with the results shown here. An important future step is to calculate the conditional activity of these mutated forms to see whether the dynamical correlations are also disrupted. For the catabolite activator protein homodimer, the conditional activity analysis indicates that the coupling exists between the cAMP-binding pocket of one monomer subunit and the DNA-binding domain of the other monomer. Coupling is absent between the other cAMP- and DNA-binding sites because the crystal structure is one of two degenerate structures in which the monomer conformations are not identical. The simulations thus sampled one of two equally stable conformational basins. It would be of interest to further investigate, via more comprehensive simulations, the interconversion between these two basins as well as the existence of other metastable basins in which both or neither of the couplings are activated.

The picture of local fluctuations at the allosteric site affecting fluctuations at the active site is also supported by acid unfolding and proton binding data on SNase showing that local fluctuations modulate ligand-binding affinities.[6] In fact, such fluctuation data can be used instead of molecular dynamics data to build conditional activity maps using experimental data. Finally, the role of rare yet long-lived conformational states in providing the disproportionate fraction of protein functionality has recently been experimentally demonstrated for a population of identical proteins,[32] in support of the finding that proteins can be in states of long as well as short waiting times between dynamical events, and that the different states can exhibit different functions.

The conditional activity is a general tool for probing dynamical correlations and can be used in future to predict functional modules and to find the dynamical connections, including those responsible for allostery, of a wide range of proteins for which such information is not readily obtained experimentally. For example, intrinsic collective dynamics is intimately linked to the catalytic function of proteins, even in the absence of the catalytic substrate. Such dynamics can encompass a subnetwork of dynamical fluctuations that is not limited to the catalytic site[33] and may be captured using the conditional activity. More broadly, in the case of signaling proteins with multiple binding partners, do different binding permutations lead to unique dynamical correlations that indicate the (in)activation of different downstream binding partners? It would also be illustrative to compare the dynamically-identified memory modules with the structurally identified autonomously folded fragments,[34] as well as the coevolutionarily identified protein sectors.[35]

Taken together, these results demonstrate that the dynamical heterogeneity within proteins gives rise to correlations in timing that can span many nanometers even if structural correlations are short-range; this is because structural side-chain correlations, unlike timing correlations, are not transitive in a noisy environment. This means that dynamical correlations may be necessary for allosteric interactions between distant side-chains in the absence of overt backbone conformational

DOI: 10.1021/jacs.5b08814
J. Am. Chem. Soc. 2016, 138, 5036−5043

change. The conditional activity reveals regions of highly non-Poissonian dynamics coalescing into functionally relevant modules as well as allosteric connections.

Outside of its application to protein dynamics, the conditional activity approach quantifies a new conceptual framework for capturing cooperativity in the time domain, and is suited to finding collective relationships in any type of collective system whose members undergo history-dependent discrete dynamics. As such, it could find wide application in the study of complex biological networks.

METHODS

Simulation and Analysis Details. The simulations were performed using the parallelized GROMACS 4.5[36,37] package using the AMBER03 force field.[38] V-sites[39] were used to coarse-grain hydrogen vibrations, and a simulation time step of 5 fs was used. Structures were saved every 200 fs. All equilibrated simulations were performed at 298.5 K and 1 atm pressure, with constant temperature and pressure maintained using the Nosé−Hoover[40,41] thermostat and Parrinello−Raman[42] barostat. Explicit water molecules were simulated using the TIP3P[43] water model.

PDZ3. Structure is taken from the pdb file 1BFE.[26] 5488 TIP3P water molecules and 4 sodium ions solvated the protein in a charge-neutral cubic box with periodic boundary conditions. Two microseconds of NPT simulations were obtained in a single run.

Lysozyme. Structure was taken from the pdb file 1AKI.[27] 12 359 TIP3P water molecules and 8 chlorine ions solvated the protein in a charge-neutral cubic box with periodic boundary conditions. Two microseconds of NPT simulations were obtained in a single run.

Catabolite Activator Protein. Structure was taken from the pdb file 1G6N,[23] and cAMP ligands were removed from the pdb file prior to simulation. 16 179 TIP3P water molecules solvated a rhombic dodecahedron box with periodic boundary conditions. Three microseconds of NPT simulations were obtained in a single run.

Visual analysis and illustrations were performed using VMD[44] and PyMOL,[45] respectively. A side-chain dihedral transition is deemed to have occurred at time t if the peak closest to the dihedral at time $t + \delta t$ is different from that at time t; the dihedral hopped from one basin to another at time t. As long as $\delta t < 20$ ps, the results are not sensitive to the choice of δt, which was chosen to be 5 ps. The statistical errors arise from the finite number of distinct transitions that occur in the simulation window, and are larger for degrees of freedom that undergo rare fluctuations. Possible error arises from the probability that a finite sampling of two independent degrees of freedom will seem to yield correlations due to random chance. For both the mutual information and the conditional activity, the error bars are found by calculating the correlation functions on same-size data sets for which the degrees of freedom are temporally decoupled from each other; the decoupling was performed by shifting the time coordinates of each transition of a dihedral angle by a random time interval chosen with uniform probability from the range [−200 ns, 200 ns]. By averaging the value of the correlation function over 100 repeats of this randomization procedure for all degrees of freedom, the expected spurious correlation due to finite sampling size were calculated and used as the error bars of the inter-residue correlation functions in Figure 3.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/jacs.5b08814.

Backbone RMSD, both in aggregate and as a function of location in catabolite activator protein, are shown in the Figure S1. The top two principal eigenvectors, projected onto the protein structure, are shown in Figure S2. The four lowest frequency (nonrotational/translational) normal modes of the protein are visualized in the movies. (PDF)

Movie S1. (MPG)
Movie S2. (MPG)
Movie S3. (MPG)
Movie S4. (MPG)

AUTHOR INFORMATION

Corresponding Author

*milo.lin@utsouthwestern.edu

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The author gratefully acknowledges the Miller Institute for Basic Research in Science at UC Berkeley for funding this work. David Chandler and Kelsey Schuster stimulated the direction of this work, and David Chandler provided valuable feedback during the research and writing phases. Gregory Bowman, Thomas Tombrello, and Susan Marqusee provided helpful comments on the manuscript.

REFERENCES

(1) Anfinsen, C. B. *Biochem. J.* **1972**, *128*, 737−749.
(2) Monod, J.; Wyman, J.; Changeux, J.-P. *J. Mol. Biol.* **1965**, *12*, 88−118.
(3) Koshland, D. E., Jr. *J. Theor. Biol.* **1962**, *2*, 75−86.
(4) Popovych, N.; Sun, S.; Ebright, R. H.; Kalodimos, C. G. *Nat. Struct. Mol. Biol.* **2006**, *13*, 831−838.
(5) Gasper, P. M.; Fuglestad, B.; Komives, E. A.; Markwick, P. R. L.; McCammon, J. A. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 21216−21222.
(6) Whitten, S. T.; E, B. G.-M.; Hilser, V. J. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 4282−4287.
(7) Motlagh, H. N.; Wrabl, J. O.; Li, J.; Hilser, V. J. *Nature* **2014**, *508*, 331−339.
(8) Dror, R. O.; Green, H. F.; Valant, C.; Borhani, D. W.; Valcourt, J. R.; Pan, A. C.; Arlow, D. H.; Canals, M.; Lane, J. R.; Rahmani, R.; Baell, J. B.; Sexton, P. M.; Christopoulos, A.; Shaw, D. E. *Nature* **2013**, *503*, 295−299.
(9) Bowman, G. R.; Geissler, P. L. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 11681−11686.
(10) Cooper, A.; Dryden, D. *Eur. Biophys. J.* **1984**, *11*, 103−109.
(11) Keys, A. S.; Hedges, L. O.; Garrahan, J. P.; Glotzer, S. C.; Chandler, D. *Phys. Rev. X* **2011**, *1*, 021013.
(12) Hedges, L. O.; Maibaum, L.; Chandler, D.; Garrahan, J. P. *J. Chem. Phys.* **2007**, *127*, 211101.
(13) Schuster, K.; Bowman, G. R.; Chandler, D. Facilitated intermittent dynamics in protein side-chains, unpublished.
(14) Cover, T. M.; Thomas, J. A. *Elements of Information Theory*, 2nd ed.; Wiley & Sons: New York, 2006.
(15) Jung, Y.; Garrahan, J. P.; Chandler, D. *J. Chem. Phys.* **2005**, *123*, 084509.
(16) Wand, A. J. *Nat. Struct. Biol.* **2001**, *8*, 926−931.
(17) Lindorff-Larsen, K.; Best, R. B.; DePristo, M. A.; Dobson, C. M.; Vendruscolo, M. *Nature* **2005**, *433*, 128−132.
(18) Igumenova, T. I.; Frederick, K. K.; Wand, A. J. *Chem. Rev.* **2006**, *106*, 1672−1699.
(19) DuBay, K. H.; Geissler, P. L. *J. Mol. Biol.* **2009**, *391*, 484−497.
(20) Bowman, G. R.; Geissler, P. L. *J. Phys. Chem. B* **2014**, *118*, 6417−6423.
(21) Tsai, C.-J.; del Sol, A.; Nussinov, R. *J. Mol. Biol.* **2008**, *378*, 1−11.
(22) DuBay, K. H.; Bothma, J. P.; Geissler, P. L. *PLoS Comput. Biol.* **2011**, *7*, e1002168.
(23) Passner, J. M.; Schultz, S. C.; Steitz, T. A. *J. Mol. Biol.* **2000**, *304*, 847−859.

DOI: 10.1021/jacs.5b08814
J. Am. Chem. Soc. 2016, 138, 5036−5043

(24) Niu, W.; Kim, Y.; Tau, G.; Heyduk, T.; Ebright, R. H. *Cell* **1996**, *87*, 1123−1134.
(25) Lange, O. F.; Grubmüller, H. *Proteins: Struct., Funct., Genet.* **2006**, *62*, 1053−1061.
(26) Doyle, D. A.; Lee, A.; Lewis, J.; Kim, E.; Sheng, M.; MacKinnon, R. *Cell* **1996**, *85*, 1067−1076.
(27) Artymiuk, P. J.; Blake, C. *J. Mol. Biol.* **1981**, *152*, 737−762.
(28) Petit, C. M.; Zhang, J.; Sapienza, P. J.; Fuentes, E. J.; Lee, A. L. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 18249−18254.
(29) Sethi, A.; Eargle, J.; Black, A. A.; Luthy-Schulten, Z. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 6620−6625.
(30) Girvan, M.; Newman, M. E. J. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 7821−7826.
(31) Yu, S.; Lee, J. C. *Biochemistry* **2004**, *43*, 4662−4669.
(32) Iversen, L.; Tu, H.-L.; Lin, W.-C.; Christensen, S. M.; Abel, S. M.; Iwig, J.; Wu, H.-J.; Gureasko, J.; Rhodes, C.; Petit, R. S.; Hansen, S. D.; Thill, P.; Yu, C.-H.; Stamou, D.; Chakraborty, A. K.; Kuriyan, J.; Groves, J. T. *Science* **2014**, *345*, 50−54.
(33) Eisenmesser, E. Z.; Millet, O.; Labeikovsky, W.; Korzhnev, D. M.; Wolf-Watz, M.; Bosco, D. A.; Skalicky, J. J.; Kay, L. E.; Kern, D. *Nature* **2005**, *438*, 117−121.
(34) Fischer, K. F.; Marqusee, S. *J. Mol. Biol.* **2000**, *302*, 701−712.
(35) Halabi, N.; Rivoire, O.; Leibler, S.; Ranganathan, R. *Cell* **2009**, *138*, 774−786.
(36) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. *J. Comput. Chem.* **2005**, *26*, 1701−1718.
(37) Pronk, S.; Pall, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. *Bioinformatics* **2013**, *29*, 845−854.
(38) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999−2012.
(39) Feenstra, K. A.; Hess, B.; Berendsen, H. *J. Comput. Chem.* **1999**, *20*, 786−798.
(40) Nosé, S. A. *Mol. Phys.* **1984**, *52*, 255−268.
(41) Hoover, W. G. *Phys. Rev. A: At., Mol., Opt. Phys.* **1985**, *31*, 1695−1697.
(42) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182.
(43) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.
(44) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33−38.
(45) Schrodinger, LLC. *The PyMOL Molecular Graphics System*, Version 1.5.0.4.

DOI: 10.1021/jacs.5b08814
J. Am. Chem. Soc. 2016, 138, 5036−5043