JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Comparing Conformational Ensembles Using the Kullback–Leibler Divergence Expansion

Christopher L. McClendon,[*,†,∥] Lan Hua,[‡] Gabriela Barreiro,[§,⊥] and Matthew P. Jacobson[‡]

[†]Graduate Group in Biophysics, [‡]Department of Pharmaceutical Chemistry, and [§]Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California 94105, United States

ABSTRACT: We present a thermodynamical approach to identify changes in macromolecular structure and dynamics in response to perturbations such as mutations or ligand binding, using an expansion of the Kullback–Leibler Divergence that connects local population shifts in torsion angles to changes in the free energy landscape of the protein. While the Kullback–Leibler Divergence is a known formula from information theory, the novelty and power of our implementation lies in its formal developments, connection to thermodynamics, statistical filtering, ease of visualization of results, and extendability by adding higher-order terms. We present a formal derivation of the Kullback–Leibler Divergence expansion and then apply our method at a first-order approximation to molecular dynamics simulations of four protein systems where ligand binding or pH titration is known to cause an effect at a distant site. Our results qualitatively agree with experimental measurements of local changes in structure or dynamics, such as NMR chemical shift perturbations and hydrogen–deuterium exchange mass spectrometry. The approach produces easy-to-analyze results with low background, and as such has the potential to become a routine analysis when molecular dynamics simulations in two or more conditions are available. Our method is implemented in the MutInf code package and is available on the SimTK website at https://simtk.org/home/mutinf.

1. INTRODUCTION

It is by now well understood that macromolecules, under biologically relevant conditions, do not adopt single conformations but display varying degrees of conformational dynamics. Equilibrium properties are thus characterized by an ensemble of conformations. Any perturbation to the system can change the energy landscape and the associated conformational ensembles; that is, both the "average" or dominant conformation and the dynamics can change. These perturbations can include environmental conditions such as the solvent or temperature, ligand binding, mutations, or post-translational modifications, and can have functional consequences. For example, mutations can modulate ligand binding in this way, leading to drug resistance, and ligand binding or post-translational modification can regulate enzyme activity. Changes in conformation and dynamics need not be confined locally but can extend across the macromolecule, leading to allostery (in the broad, modern use of the word). Molecular dynamics simulations and related computational methods provide practical ways to generate macromolecular ensembles, although all such methods are limited by incomplete sampling. Many analysis methods are commonly employed to characterize the resulting ensembles, and to compare ensembles. Although routine, such analyses are fundamentally challenging because of the large number of degrees of freedom and the complexity of the conformational ensembles, which are not always well approximated by fluctuations around an "average" structure. Approaches to compare molecular conformational ensembles can focus on global phenomena or localized phenomena. Approaches for capturing global differences between conformational ensembles typically reduce the dimensionality by discretizing conformational space over a subset of degrees of freedom (i.e., Cα atoms) into rapidly converting "microstates" and slowly converting "macrostates",[1] by changing basis into a subset of the most significant collective coordinates by performing some variant of principal coordinates analysis.[2−4] Approaches for capturing localized differences between conformational ensembles typically focus on average structural changes, average flexibility changes, contact maps,[5] or correlated motions.[6,7] In this work, we describe a computational approach for comparing conformational ensembles, based on the Kullback–Leibler Divergence from information theory, which captures both conformational changes and changes in entropy/dynamics. The results are thermodynamically meaningful, easy-to-visualize, and filtered for statistical significance so that significant perturbations are easy to identify. The approach is well suited to such tasks as identifying, in an unbiased way, whether perturbations such as ligand or protein binding or post-translational modification alter conformational ensembles at distant sites, and whether two different ligands binding to the same site cause similar or different effects.

2. OVERVIEW OF METHOD

We analyze residues' conformational distributions in torsion space, as torsions provide an apt local description of biologically relevant functional motions and do not have frame-fitting issues inherent to Cartesian analysis.[3] Especially for protein side chains, the concept of "average" positions is of limited use, as their distributions are often multimodal in torsional or Cartesian space. To quantify "population shifts" in residues' conformational distributions, we use the Kullback–Leibler Divergence, a measure of the free energy difference between

Received: January 6, 2012
Published: April 13, 2012

two equilibrium ensembles, where one ensemble is the "reference" ensemble and the other is the "perturbed" ensemble.[8,9] The Kullback–Leibler Divergence or relative entropy is a fundamental quantity in information theory, and its differential version is given by:

$$\mathrm{KL}(x_1, ..., x_m \| x_1{}^*, ..., x_m{}^*) = \int J(x_1, ..., x_m)\rho(x_1, ..., x_m)\ln \frac{\rho(x_1, ..., x_m)}{\rho^*(x_1{}^*, ..., x_m{}^*)}\mathrm{d}x_1, ..., \mathrm{d}x_m \quad (1)$$

where ρ^* is the probability density function (p.d.f.) of the reference ensemble, ρ is the p.d.f. of the perturbed ensemble, and J indicates the Jacobian determinant. Torsion angles (with fixed bond lengths and angles) or orthogonal Cartesian basis sets have a Jacobian of unity, facilitating analysis. The Kullback–Leibler Divergence was previously derived to second order for a harmonic Hamiltonian and applied to normal-modes models of proteins.[10] It was applied to trypsinogen to not only refine a normal-mode model against atomistic simulation data, but also to quantify coupling between trypsinogen's active and regulatory sites. To our knowledge, this study was the first to apply the Kullback–Leibler Divergence to studying allostery. In a different study, this measure was applied to identify functional sites in a large test set of proteins.[11] This approach to identify functional sites was based on the observation that functional sites tended to colocalize with surface sites where artificial perturbations caused a large change in the total Kullback–Leibler Divergence for the protein. The Kullback–Leibler Divergence was also applied at second order in a perturbational formulation of principal components analysis (PCA) to identify effective perturbations that contribute to differences between conformational ensembles.[3] In PCA of Cα atoms, a common practice in molecular dynamics simulations, these perturbation functions are identity operators on the Cartesian coordinates. The eigenvalues of the perturbation functions' covariance matrix (or that of the Cα coordinates in typical applications of PCA) are related to the Kullback–Leibler Divergence between ensembles. To calculate the Kullback–Leibler Divergence for macromolecular conformational ensembles containing many degrees of freedom, we propose an expansion over increasing numbers of degrees of freedom. We derive a novel expansion of the K–L divergence over single degrees of freedom, pairs of degrees of freedom, etc., utilizing the Generalized Kirkwood Superposition Approximation (GKSA), which has been previously used by Matsuda[12] and by Killian et al. for a configurational entropy expansion.[13] The most immediate application is the use of first-order terms to calculate the Kullback–Leibler Divergence for protein residues from sums of the Kullback–Leibler Divergences of their constituent torsions, which could be readily refined by use of second-order terms within residues. We expect that second- and higher-order terms will also be useful in future applications. Importantly, our expansion connects such "local" Kullback–Leibler Divergences to the global Kullback–Leibler Divergence for the conformational ensemble, which has connections to the free energy; other measures of comparison such as rms deviation or chi-squared analysis lack this strong connection to thermodynamics. Our method scales linearly with the number of residues in the protein (neglecting inter-residue second-order and higher-order terms) and is thus applicable to large macromolecules and complexes. The novelty of our work lies in: (1) providing a thermodynamics-based comparison between conformational ensembles that accounts for both changes in structure and in flexibilities, in contrast with commonly used methods such as root-mean-squared deviation (rmsd) or root-mean-squared fluctuation (RMSF or B-factor analysis); (2) deriving an expansion that prescribes a way to compare distributions of multiple degrees of freedom (e.g., the multiple torsions of a protein residue or those of a group of residues), and a systematic way to improve the accuracy of such comparisons; and (3) the introduction of a useful quantity, which we call the "mutual divergence", analogous to mutual information except that it uses relative entropy instead of entropy, and its higher-order analogue. In the numerical implementation, we also provide a discretization correction to the Kullback–Leibler Divergence, and use bootstrap resampling on the Kullback–Leibler Divergence for statistical filtering and correction of sampling bias.

2.1. Marginal Probability Distributions and the Generalized Kirkwood Superposition Approximation.

A protein's geometry is most commonly described in Cartesian coordinates or in internal bond-angle-torsion (BAT) coordinates. We use BAT coordinates and focus our analysis on ϕ, ψ, and χ torsion angles, as these are the most important to describe motions of biophysical relevance. The distribution of the m torsion angles $(x_1,...,x_m)$ of a protein's "perturbed" equilibrium conformational ensemble (perturbed by mutation, ligand binding, post-translational modification, etc.) gives rise to a probability distribution $\rho(x_1,...,x_m)$ over m degrees of freedom; these are compared to a "reference" conformational ensemble having probability distribution ρ^*. The number of snapshots of a protein's geometry required to adequately approximate this m-dimensional probability distribution function (p.d.f.) grows exponentially with increasing m. For this reason, we wish to approximate the m-dimensional p.d.f. using marginal distributions of $\rho(x_1,...,x_m)$ involving only one and two variables. Such marginal distributions of order n are defined as follows:

$$\rho_1(x_j) = \int J(x_1, ..., x_m)\rho(x_1, ..., x_m) \prod_{i\neq j} \mathrm{d}x_i = \rho_{1,\mathbf{s}=\{j\}} \quad (2)$$

$$\rho_2(x_j, x_k) = \int J(x_1, ..., x_m)\rho(x_1, ..., x_m) \prod_{i\neq j,k} \mathrm{d}x_i = \rho_{2,\mathbf{s}=\{j,k\}} \quad (3)$$

$$\rho_n(x_{j1}, ..., x_{jn}) = \int J(x_1, ..., x_m)\rho(x_1, ..., x_m) \prod_{i\neq \{j_n\}} \mathrm{d}x_i = \rho_{n,\mathbf{s}=\{x_1,...,x_{n-1}\}} \quad (4)$$

where $\mathbf{s}$ denotes a set of degrees of freedom. In what follows, the subscript of probability densities $\rho_{n,k}$ will either have one index indicating the number of degrees of freedom and an argument list, or be expressed in shortened notation using two indices: the first, n, indicating the number of degrees of freedom in the probability density function, and the second, $\{k\}$, indicating a set of indices of degrees of freedom comprising the p.d.f. The Generalized Kirkwood Superposition Approximation (GKSA) is of key importance for the foundation of the present work. The GKSA at order $m - 1$ approximates a probability distribution with m degrees of freedom using lower-order probability density functions consisting of a subset of the degrees of freedom, up to $m - 1$ degrees of freedom for an order $m - 1$ GKSA, and is perhaps easiest to express in log form:

$$\ln \hat{\rho}_m^{m-1}(x_1, ..., x_m) = \sum_{n=1}^{m-1} (-1)^{m-n+1} \ln \prod_{\mathbf{k}}^{C_n^m} \rho_{n,\mathbf{k}} \tag{5}$$

where C_n^m indicates all $\binom{m}{n}$ combinations of nth-order marginal probability density functions of ρ, and $\hat{\rho}^{m-1}$ indicates the order $m - 1$ GKSA approximation of ρ. As it has been noted that the terms in this superposition are not appropriately normalized p.d.f.'s except for the first-order terms,[14] it is not clear whether a GKSA-based expansion of the total Kullback–Leibler Divergence would be expected to give quantitative measures of the total free energy cost of remodeling the conformational distribution or free energy landscape of a macromolecule. Nonetheless, the success of the configurational entropy expansion and its variants in computing configurational entropies suggests that the total Kullback–Leibler Divergence under this approximation may still be of use beyond the relative values of its terms applied in the Results.

3. METHODS

3.1. Kullback–Leibler Divergence Expansion for Three Variables. To motivate the expansion of the Kullback–Leibler Divergence, consider a probability distribution $\rho(x_1,...,x_m) = \rho(\phi,\psi,\chi)$ that is a function of a set τ of three variables, $\tau = \{\phi,\psi,\chi\}$. Suppose, for example, that these three variables denote the backbone and first side chain torsion angles of an amino acid in a peptide or protein. The Kirkwood expansion for ρ is then:

$$\rho_3(\phi, \psi, \chi) \approx \frac{\rho_2(\phi, \psi)\rho_2(\phi, \chi)\rho_2(\psi, \chi)}{\rho_1(\phi)\rho_1(\psi)\rho_1(\chi)} = \frac{\prod_{\mathbf{g}}^{C_2^3} \rho_{2,\mathbf{g}}}{\prod_{k}^{C_1^3} \rho_{1,\mathbf{k}}} \tag{6}$$

where the notation C_p^q denotes all q-choose-p combinations of order-p marginal distributions, and **g** and **k** denote two-member and one-member sets of degrees of freedom comprising a particular combination of these C_p^q order-p marginals. Consider probability distributions ρ and ρ^* over m degrees of freedom. Continuing with our example, inserting the Kirkwood expansion for ρ and ρ^* into the equation above yields:

$$\mathrm{KL}(x_1, ..., x_m \| x_1^*, ..., x_m^*) = \int \rho(x_1, ..., x_m) \ln \frac{\left(\prod_{\mathbf{g}}^{C_2^3} \rho_{2,\mathbf{g}}\right) / \left(\prod_{\mathbf{k}}^{C_1^3} \rho_1\right)}{\left(\prod_{\mathbf{g}}^{C_2^3} \rho^*_{2,\mathbf{g}}\right) / \left(\prod_{\mathbf{k}}^{C_1^3} \rho^*_{1,\mathbf{k}}\right)} \mathrm{d}x_1, ..., \mathrm{d}x_m \tag{7}$$

Converting the log of a product into a sum of logs:

$$\mathrm{KL}(x_1, ..., x_m \| x_1^*, ..., x_m^*) = \int \rho(x_1, ..., x_m) \left(\sum_{\mathbf{g}}^{C_2^3} \ln \frac{\rho_{2,\mathbf{g}}}{\rho_{2,\mathbf{g}}^*} - \sum_{\mathbf{k}}^{C_1^3} \ln \frac{\rho_{1,\mathbf{k}}}{\rho_{1,\mathbf{k}}^*} \right) \mathrm{d}x_1, ..., \mathrm{d}x_m \tag{8}$$

Because of linearity:

$$\begin{aligned} &\mathrm{KL}(x_1, ..., x_m \| x_1^*, ..., x_m^*) \\ &= \sum_{\mathbf{g}}^{C_2^3} \int \rho(x_1, ..., x_m) \ln \frac{\rho_{2,\mathbf{g}}}{\rho_{2,\mathbf{g}}^*} \mathrm{d}x_1, ..., \mathrm{d}x_m \\ &\quad - \sum_{\mathbf{k}}^{C_1^3} \int \rho(x_1, ..., x_m) \ln \frac{\rho_{1,k}}{\rho_{1,k}^*} \mathrm{d}x_1, ..., \mathrm{d}x_m \end{aligned} \tag{9}$$

For each of these sums of C_n^m combinations of log terms, we can integrate out the $m - n$ degrees of freedom that are not part of each log term, and define $\mathrm{d}\tau^n$ as the differential volume element over the remaining n variables in each term.

$$\begin{aligned} &\mathrm{KL}(x_1, ..., x_m \| x_1^*, ..., x_m^*) \\ &= \left(\sum_{\mathbf{g}}^{C_2^3} \int \rho_{2,\mathbf{g}} \ln \frac{\rho_{2,\mathbf{g}}}{\rho_{2,\mathbf{g}}^*} \mathrm{d}\tau^2 \right) - \left(\sum_{\mathbf{k}}^{C_1^3} \int \rho_{1,\mathbf{k}} \ln \frac{\rho_{1,\mathbf{k}}}{\rho_{1,\mathbf{k}}^*} \mathrm{d}\tau^1 \right) \end{aligned} \tag{10}$$

Expanding, we see that this is merely the sum of Kullback–Leibler Divergences of pairwise p.d.f.'s (with respect to their equilibrium values) minus the Kullback–Leibler Divergences of individual p.d.f.'s:

$$\begin{aligned} &\mathrm{KL}(x_1, ..., x_m \| x_1^*, ..., x_m^*) \\ &= -\int \rho_1(\phi) \ln \frac{\rho_2(\phi)}{\rho_2^*(\phi)} \mathrm{d}\phi - \int \rho_1(\psi) \ln \frac{\rho_2(\psi)}{\rho_2^*(\psi)} \mathrm{d}\psi \\ &\quad - \int \rho_1(\chi) \ln \frac{\rho_2(\chi)}{\rho_2^*(\chi)} \mathrm{d}\chi \\ &\quad + \int \rho_2(\phi, \psi) \ln \frac{\rho_2(\phi, \psi)}{\rho_2^*(\phi, \psi)} \mathrm{d}\phi \mathrm{d}\psi \\ &\quad + \int \rho_2(\phi, \chi) \ln \frac{\rho_2(\phi, \chi)}{\rho_2^*(\phi, \chi)} \mathrm{d}\phi \mathrm{d}\chi \\ &\quad + \int \rho_2(\psi, \chi) \ln \frac{\rho_2(\psi, \chi)}{\rho_2^*(\psi, \chi)} \mathrm{d}\psi \mathrm{d}\chi \end{aligned} \tag{11}$$

3.2. General Derivation of Kullback–Leibler Divergence Expansion. Now that we have illustrated the Kullback–Leibler Divergence Expansion for three degrees of freedom, we next provide a general derivation of the expansion to m degrees of freedom, following similar procedures used in the entropy expansion in Killian et al.[13] and in Matsuda.[12] Applying the GKSA approximation to ρ and ρ^* inside the logarithm of eq 1:

$$\begin{aligned} \mathrm{KL}^{m-1} = &\int J(x_1, ..., x_m) \rho_m(x_1, ..., x_m) \sum_{n=1}^{m-1} (-1)^{m-n+1} \\ &\times \ln \prod_{\mathbf{k}}^{C_n^m} \frac{\rho_{n,\mathbf{k}}}{\rho^*_{n,\mathbf{k}}} \mathrm{d}x_1, ..., \mathrm{d}x_m \end{aligned} \tag{12}$$

The superscript above KL denotes the order of the approximation to the Kullback–Leibler Divergence. Again, C_n^m indicates all $\binom{m}{n}$ combinations of nth-order marginal probability density functions of ρ, and $\hat{\rho}^{m-1}$ will later indicate the order $m - 1$ GKSA approximation of ρ. As before, **k** denotes n-member sets of degrees of freedom comprising a particular combination of these C_n^m order-n marginals.

Converting the log of the product into a sum over logs and taking the sum outside the integral:

$$\mathrm{KL}^{m-1} = \sum_{n=1}^{m-1} (-1)^{m-n+1} \sum_{\mathbf{k}}^{C_n^m} \int J(x_1, ..., x_n)\rho_m(x_1, ..., x_m) \times \ln \frac{\rho_{n,\mathbf{k}}}{\rho_{n,\mathbf{k}}{}^*} \mathrm{d}x_1, ..., \mathrm{d}x_m \tag{13}$$

We then integrate over the $m - n$ dimensions that are independent of the log terms; these each integrate to unity. Next, define $\mathrm{d}\tau^n$ as the differential element of volume corresponding to the n dimensions that remain (including the remaining portions of the Jacobian determinant):

$$\mathrm{KL}^{m-1} = \sum_{n=1}^{m-1} (-1)^{m-n+1} \left(\sum_{\mathbf{k}}^{C_n^m} \int \rho_{n,\mathbf{k}} \ln \frac{\rho_{n,\mathbf{k}}}{\rho_{n,\mathbf{k}}{}^*} \mathrm{d}\tau^n \right) \tag{14}$$

As the term in braces in just an nth order joint K−L divergence associated with each subset $\mathbf{k}$ of n degrees of freedom chosen from $(x_1,..,x_m)$, which we denote as $\mathrm{KL}_{n,\mathbf{k}}$, this simplifies to:

$$\mathrm{KL}^{m-1} = \sum_{n=1}^{m-1} (-1)^{m-n+1} \left(\sum_{\mathbf{k}}^{C_n^m} \mathrm{KL}_{n,\mathbf{k}} \right) \tag{15}$$

To calculate the requisite integrals, we can partition m-dimensional continuous torsional space into a discrete space of histogram bins. Each degree of freedom's marginal p.d.f. is discretized into histogram bin probabilities p_i (with reference counts p_i*), and joint histograms for marginal p.d.f.'s involving pairs of degrees of freedom are given by bin probabilities p_{ij} (with reference counts p_{ij}*). These probabilities each must sum to unity: $\sum_i p_i = 1$, $\sum_{ij} p_{ij} = 1$. This partitioning leads to the following expansion over contributions from single degrees of freedom, pairs, triples, etc., for m degrees of freedom:

$$\mathrm{KL} = (-1)^m \sum_{\mathbf{k}}^{C_1^m} \sum_i^{n\text{bins}} p_i \ln \frac{p_i}{p_i^*} + (-1)^{m-1} \sum_{\mathbf{g}}^{C_2^m} \sum_{ij}^{n\text{bins}^2} p_{ij} \ln \frac{p_{ij}}{p_{ij}^*} + ... \tag{16}$$

Note that the signs of the terms depend on the number of terms: this is not ideal for an expansion. Thus, we need to introduce a term for the Kullback−Leibler Divergence that will play the same role as the mutual information in the entropy expansion of Matsuda.[12] We call this the Mutual Divergence, M, between two degrees of freedom, with marginal p.d.f.'s specified by p_i and p_j and joint histogram p_{ij} in one ensemble (the "target" ensemble), and with marginal p.d.f.'s specified by p_i and p_j and joint histogram p_{ij} in another ensemble (the "reference" ensemble):

$$M_2 = \sum_i p_i \ln \frac{p_i}{p_i^*} + \sum_j p_j \ln \frac{p_j}{p_j^*} - \sum_i \sum_j p_{ij} \ln \frac{p_{ij}}{p_{ij}^*} \tag{17}$$

This can be equivalently expressed by combining terms into a single argument in the logarithm:

$$M_2 = \sum_i \sum_j p_{ij} \ln \frac{p_i p_j p_{ij}^*}{p_i^* p_j^* p_{ij}} \tag{18}$$

Here, the sums over i, j, and ij refer to one- and two-dimensional p.d.f.'s from a given pair of degrees of freedom. Alternatively, we can view this mutual divergence M_2 as a cross-information minus the mutual information of the target state:

$$M_2 = \sum_i \sum_j p_{ij} \ln \frac{p_{ij}^*}{p_i^* p_j^*} - I(p_i, p_j) \tag{19}$$

where $I(p_i^*, p_j^*)$ indicates the mutual information between these p.d.f.'s. Mutual divergence can be generalized to higher order to provide a mutual divergence between n degrees of freedom:

$$M_n(x_1, ..., x_n \| x_1^*, ..., x_n^*) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{i_1 < \cdots < i_k} \mathrm{KL}(x_{i_1}, ..., x_{i_k} \| x_{i_1}^*, ..., x_{i_k}^*) \tag{20}$$

Moreover, the mutual divergence satisfies a recursion relation analogous to Matsuda's recursion relation for higher-order mutual information:[12]

$$\begin{aligned} M_n(x_1, ..., x_n \| x_1^*, ..., x_n^*) &= M_{n-1}(x_1, ..., x_{n-2}, x_{n-1} \| x_1^*, ..., x_{n-2}^*, x_{n-1}^*) \\ &\quad + M_{n-1}(x_1, ..., x_{n-2}, x_n \| x_1^*, ..., x_{n-2}^*, x_n^*) \\ &\quad - M_{n-1}(x_1, ..., x_{n-2}, x_{n-1}x_n \| x_1^*, ..., x_{n-2}^*, x_{n-1}^* x_n^*) \end{aligned} \tag{21}$$

Here, $x_{n-1}x_n$ indicates the joint distribution of these n degrees of freedom, as in the third term of eq 17. In terms of probability densities, the higher-order mutual divergence between n degrees of freedom is given by:

$$M_n(x_1, ..., x_n \| x_1^*, ..., x_n^*) = (-1)^n \sum_{x_1, ..., x_n} p_n(x_1, ..., x_n) \times \ln\left(\frac{p_n^*(x_1, ..., x_n)\hat{p}_{n-1}(x_1, ..., x_n)}{\hat{p}_{n-1}^*(x_1, ..., x_n) p_n(x_1, ..., x_n)} \right) \tag{22}$$

where p_n is the target distribution for these n degrees of freedom, p_n* is the reference distribution, and $\hat{p}_{n-1}$ and $\hat{p}_{n-1}$* are their Generalized Kirkwood Superposition Approximations, which consist of up to order $m - 1$ probability densities. It is worth noting that the argument of the log is inverted with respect to an analogous expression for the mutual information because there is a sign change that comes from the fact that entropies are based on $-p \ln p$ terms, while Kullback−Leibler Divergences are based on $p \ln(p/p^*)$ terms. Applying this relation to the Kullback−Leibler Divergence, we obtain the desired expansion over m degrees of freedom:

$$\mathrm{KL} = \sum_{n=1}^{m} \sum_i^{n\text{bins}} p_i \ln \frac{p_i}{p_i^*} - \sum_{n=1}^{m} \sum_{n' \neq n} \sum_{ij}^{n\text{bins}^2} p_{ij} \ln \frac{p_i p_j p_{ij}^*}{p_i^* p_j^* p_{ij}} + ... \tag{23}$$

Although this expansion and the previous expansion, eq 16, agree when all terms are present, in practice this expansion in eq 23 is far more useful as it provides a well-defined way to truncate the expansion at a given complexity.

3.3. Local Kullback−Leibler Divergence.

We are interested in population shifts caused by perturbations that reflect subtle changes in structure and/or dynamics in particular protein residues. We can visualize these most readily using the first-order terms from our expansion. Consider the terms in the Kullback−Leibler Divergence arising from a particular degree of freedom. These we will denote the "local" Kullback−Leibler Divergence and provide an information-theoretic, quantitative measure of the extent to which the p.d.f. for a given degree of freedom deviates from the equilibrium p.d.f. This quantifies changes in probability density with fewer assumptions and a better connection to thermodynamics than the more familiar chi-squared statistic.

$$\mathrm{KL}_1 = \sum_{i}^{n\mathrm{bins}} p_i \ln \frac{p_i}{p_i^*} \tag{24}$$

To calculate the local Kullback−Leibler Divergence for a single protein residue, we simply sum the Kullback−Leibler Divergences between the reference and target ensemble for each of the residue's ϕ, ψ, and χ torsion angles:

$$\mathrm{KL}_{\mathrm{res}_n} = \sum_{\phi,\psi,\chi'\mathrm{s}} \sum_{i}^{n\mathrm{bins}} p_i \ln \frac{p_i}{p_i^*} \tag{25}$$

While this expression is very similar to the well-known $p \ln p$ expression for entropy, with the nonuniform reference state p_i^* making this the relative entropy, it is thermodynamically distinct; it is a measure of dissimilarity of two probability density functions, rather than the disorder of a particular probability density function. While presently we focus on applications of this first-order term, which has been used to compare Markov models of conformational ensembles[1] from molecular dynamics simulations but has not been widely applied on a per-residue level, the full derivation presented here establishes a systematic approach to improve our method. At the per-residue level or at the groups-of-residues level, we could improve our method by considering pairs of torsions within a residue or set of residues, etc. Furthermore, application of our method at the pairs-of-residues level or at higher order could identify changes in correlated motions, although these require substantially more sampling than first-order terms.[15] This could be a promising direction for future research, but is beyond the scope of the present work. Here, however, we focus on first-order terms, as these are most readily, rapidly, and robustly calculated, and the computational cost scales linearly with system size.

3.4. Statistical Corrections to the Kullback−Leibler Divergence. If the "target" ensemble is the same as the equilibrium ensemble, the Kullback−Leibler Divergence will be zero. However, in practice, when applied to ensembles generated by methods such as molecular dynamics, this is not often the case due to sample variability. To improve the signal-to-noise ratio in our calculation of the Kullback−Leibler Divergence, and thereby distinguish meaningful differences between conformational ensembles from artifactual population shifts due to sample variability, we calculate the K−L divergence expected from sample variability in the "reference" ensemble and use it for a significance test and to correct the calculated values. To generate a realistic measure of sample variability, we use a statistical bootstrapping approach. We split the full reference ensembles into *n*sims blocks (usually corresponding to clones of the same system with different random number seeds, or large continuous blocks from long simulations), and take half of the blocks at a time as a surrogate target ensemble and the complementary half as a surrogate reference ensemble. We aggregate the counts for the torsions to construct probability distributions and calculate the K−L divergence between all combinations of surrogate distributions. Any nonzero average K−L divergence between these distributions is a measure of average bias that we can later subtract from the total K−L divergence between the full "reference" ensemble and the full "target" ensemble, when it is significant. The K−L divergence under the null hypothesis that the average K−L divergence is no greater than that expected from sample variability in the reference ensemble is then given by:

$$\mathrm{KL}_1^{H_0} = \binom{n\mathrm{sims}}{n\mathrm{sims}/2}^{-1} \sum_{\mathrm{blocks}}^{n\mathrm{sims}/2} \sum_{i}^{n\mathrm{bins}} p_i \ln \frac{p_i^{\mathrm{S}}}{p_i^{\mathrm{S^C}}} \tag{26}$$

where S denotes subsamples and $\mathrm{S^C}$ are their complements. To test for statistical significance of the obseved Kullback−Leibler Divergence, we use the distribution of these surrogate Kullback−Leibler Divergence values to obtain a *p*-value for the null hypothesis that the average Kullback−Leibler Divergence is no greater than that expected from sample variability in the reference ensemble. If this *p*-value for a particular torsion is less than the significance level (in this case, set at a permissive $\alpha = 0.1$), then the Kullback−Leibler Divergence is set to zero; if not, then the average Kullback−Leibler Divergence between the surrogate distributions described above is subtracted from the total, in a manner similar to corrections to mutual information:[7,16]

$$\hat{\mathrm{KL}}_1 = \mathrm{KL}_1 - \mathrm{KL}_1^{H_0} \tag{27}$$

3.5. Truncation of Kullback−Leibler Divergence. Given the expansion in eq 23, one may wonder why truncation at a particular order might be appropriate, especially as the number of terms at each order increases combinatorially before contracting toward the tail of the expansion. In the analogous configurational entropy expansion,[13] small-molecule systems achieved remarkable agreement with entropies from rigorous free energy calculations by only including first- and second-order terms in the expansion, with the highly correlated cyclohexane requiring up to third-order terms. We note that the pairwise mutual divergence between two degrees of freedom is less than or equal to the sum of the corresponding first-order Kullback−Leibler Divergence terms. Thus, for the mutual divergence to be significantly greater than zero, at least one of the constituent degrees of freedom must be statistically significant. It is important to note that higher-order terms in eq 23 capture only changes in distributions missed by lower-order terms. For example, the mutual divergence captures population shifts in pairs of degrees of freedom that are missed by the first-order Kullback−Leibler Divergence. The key parameter governing the maximal order needed for convergence of the expansion is the maximum number of coupled independent components or modes (i.e., effective dimensionality) in the system. A recent study used a novel approach to partition molecular dynamics trajectories into independent subspaces of coupled modes[17] and found a block-like pattern where groups of pairwise correlated modes had minimal couplings with other blocks of correlated modes in a 100 ns simulation of lysozyme. Specifically, there was a maximum of six modes per block, with most blocks only containing a few modes. Thus, the maximum number of coupled independent components in this study was six, so the Kullback−Leibler Divgence expansion should only require terms up to sixth order. Even though the number of terms at each order might increase, the sparsity of the matrix of mode couplings at second order suggests that a lower fraction of higher-order terms would have significant values. Other studies have also taken advantage of the sparsity of second-order couplings to more efficiently diagonalize the Hamiltonian for the protein.[18,19] To obtain better convergence of the Kullback–Leibler Divergence expansion, we could take a subset of the terms along a minimal spanning tree (treating terms as nodes), as in the MIST approach.[20] Importantly, MIST avoids the combinatorial explosion in a number of terms at higher orders. Practically, higher-order mutual divergences will require exponentially more data points sampled to give a robust estimate, as the volume of space increases exponentially with the number of degrees of freedom. Currently, only calculations up to third order might be practical with microseconds of simulation data.[15] Neglecting high-order terms should not affect our qualitative interpretation of the pattern of local, per-residue Kullback–Leibler Divergences because there are only a small number of torsions per residue.

3.6. Jensen–Shannon Divergence. The Jensen–Shannon Divergence is a slight variation of the Kullback–Leibler Divergence and has the added benefit of treating both "reference" and "target" ensembles symmetrically, albeit at a cost of possibly providing lower signal-to-noise due to averaging (see eq 28). Furthermore, the Jensen–Shannon Divergence is related to thermodynamic length, an asymptotic bound on energy dissipated in a finite-time transformation from one state to another.[21] Because the Kullback–Leibler Divergence expansion is general for any "reference" distribution, we can take the new reference distribution to be merely the superposition of the former "reference" distributions and calculate the Jensen–Shannon divergence as the mean of the Kullback–Leibler Divergences between either ensemble and this new reference distribution:

$$\begin{aligned}&\mathrm{JS}(x_1, ..., x_m||y_1, ..., y_m)\\&\quad= \frac{1}{2}\mathrm{KL}\big(x_1, ..., x_m\big|\big|(x_1, ..., x_m) + (y_1, ..., y_m)\big)\\&\qquad+ \frac{1}{2}\mathrm{KL}\big(y_1, ..., y_m\big|\big|(x_1, ..., x_m) + (y_1, ..., y_m)\big)\end{aligned} \tag{28}$$

To apply the same statistical test and filtering as above, we take the distribution of the Jensen–Shannon Divergence under the null hypothesis as the average of the null hypothesis distributions of the Jensen–Shannon Divergences within the separate "reference" and "target" ensembles. We construct the null hypothesis distribution in this way because we want the null hypothesis to only be a function of variation within the "reference" and "target" ensembles, and not depend on their superposition, which comes into play in computing the observed Jensen–Shannon Divergence between the ensembles in eq 28.

3.7. Molecular Dynamics Simulations. We illustrate the K–L method using examples of previously published molecular dynamics studies on human interleukin-2[7] and talin[22] and new molecular dynamics trajectories on a kinase, PDK1. These examples highlight the role of dynamics in protein function, particularly allostery. For the new molecular dynamics simulations of PDK1, we prepared the protein and ligand with Maestro's Protein Preparation Wizard (Schrodinger, 2009), with protonation states of histidine and Asn/Gln flips assigned by ProtAssign (Schrodinger, 2009) in the preparation wizard. Each model was solvated in SPC water[23] in a cubic simulation box, and $\mathrm{Na^+}$ and $\mathrm{Cl^-}$ ions were added to neutralize the system, and then an additional 0.1 M NaCl was added. The full simulation system was energy-minimized using Desmond[24] in two stages: (1) all protein and ligand atoms restrained with a force constant of 50.0 kcal/mol/Å^{-2}, and (2) no restraints. Minimizations were performed with no less than 10 steps of steepest descent minimization followed by L-BFGS optimization after a gradient of 50.0 kcal mol^{-1} Å^{-1} is reached up to a total of 2000 steps or a gradient of 50.0 kcal mol^{-1} Å^{-1} in step 1 or 5.0 kcal mol^{-1} Å^{-1} in step 2. After full minimization of the system, an equilibration was performed. First, the systems were run at constant temperature and volume at a temperature of 10 K for 12 ps using the Berendsen thermostat[25] with a relaxation time of 0.1 ps, half-sized timesteps (see below), and all protein and ligand atoms restrained with a force constant of 50.0 kcal/mol/Å^{-2}, and velocities randomized every 1.0 ps. Subsequently, molecular dynamics at constant temperature and pressure at 10 K was performed using the Berendsen thermostat and barostat,[25] with all protein and ligand atoms restrained, time constants of 0.1 and 50.0 ps for the thermostat and barostat, respectively, and velocities randomized every 1.0 ps. Next, molecular dynamics at the target temperature and pressure of 300 K and 1 atm were performed using the Berendsen thermostat and barostat with all protein and ligand atoms restrained for 12 ps using time constants of 0.1 and 50.0 ps for the thermostat and barostat, respectively, randomizing velocities every 1.0 ps, and finally for another 24 ps without restraints and with time constants of 0.1 and 2.0 ps for the thermostat and barostat, respectively. Production runs of 10 ns were performed on each system using the Martyna–Tobias–Klein integrator[26] with a reference temperature of 300 K and a reference pressure of 1 atm. Snapshots were output every 1.002 ps. The thermostat used an equilibrium temperature of 300 K, a relaxation time of 1 ps, chain length of 2, and update frequency of two steps for the system and for the barostat. The barostat featured a relaxation time of 2 ps, a reference pressure of 1 atm, isotropic coupling, and a compressibility of 4.5×10^{-5} bar^{-1}. Both the equilibration and the production molecular dynamics simulations were performed with all bonds involving hydrogens constrained, a 2 fs time step for the bonded and short-range nonbonded interactions, and updating of long-range nonbonded interactions every 6 fs using the RESPA multiple time step approach.[27] For the first NVT equilibration step, 1 and 3 fs timesteps were used, respectively. Short-range Coulombic and van der Waals nonbonded interactions were cutoff at 9.0 Å, and long-range electrostatics were computed using the smooth particle mesh Ewald method. Pairlists were constructed using a distance of 10.0 Å and a migration interval of 12 fs.

4. RESULTS

4.1. An Allosteric Small Molecule Activator of PDK1. PDK1 is a member of the AGC family of kinases, which includes protein kinases A (PKA), B (AKT), and C (multiple isozymes). In recent years, small molecules have been discovered that bind outside the active site and promote or inhibit activity. Precisely how these small moleules alter PDK1's activity is not known. The mechanism of one previously reported noncovalent small-molecule activator, PS48, was studied using hydrogen–deuterium exchange mass spectrometry experiments to determine which peptide regions of the kinase have amide protons that are protected from exchange with solvent deuterons.[28] In these experiments, amide protons both near the binding site and distant from the binding site (Figure 1) showed protection from solvent exchange, indicating more stable backbone hydrogen bonds and hence reduced flexibility. Interestingly, some of these protected regions include the DFG-loop (cyan) and activation loop (teal), whose proper positioning is essential for activity. Mutation of Thr226, adjacent to the DFG sequence, to alanine abolishes the ability of PS48 to activate the kinase. We used our Kullback–Leibler Divergence method to investigate the allosteric activation mechanism. We performed a series of 10 ns molecular dynamics simulations on PDK1 with and without PS48 bound (PDB: 3HRF), saving the conformations every 1 ps. We then calculated the first-order Kullback–Leibler Divergence between apo and PS48-bound conformational ensembles from the MD simulations. The results indicate that PS48 binding caused significant population shifts in the torsion angles of residues around the compound's binding site: the αC-helix, the β strands 145–149 and 154–159, and the αB-helix. Furthermore, there were significant population shifts in torsion angle populations distant from the PS48 binding site, for example, in the activation loop, the F-helix, and the "G" in the DFG-loop. Although the time scale of the simulations is short relative to the time scale probed by the hydrogen–deuterium exchange experiments, the local Kullback–Leibler Divergence values and hydrogen–deuterium exchange results show compound-induced changes in the regions protected from H/D exchange. The major discrepancy is that the experiments did not show substantial protection for C-lobe residues outside of the DFG motif and activation loop (possibly because backbone amides here are largely protected within stable α-helices), whereas in the simulations a number of these residues experienced significant population shifts upon ligand binding. Nonetheless, these results serve as a powerful demonstration of how our method can identify potential allosteric effects of ligand binding or mutation.

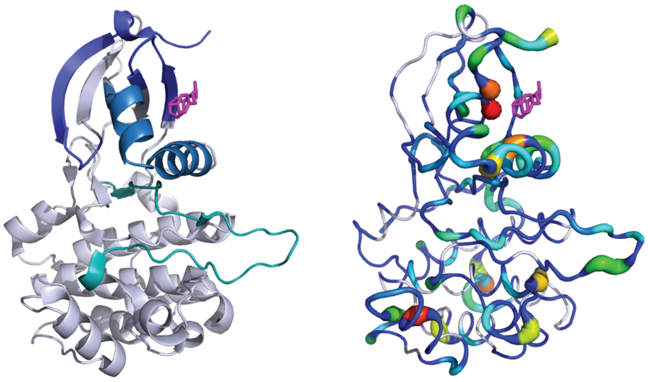

Figure 1. Kullback–Leibler Divergence highlights PDK1 regions that show protection in hydrogen–deuterium exchange experiments upon addition of an allosteric small-molecule activator. (Left) A small-molecule activator of PDK1 was previously shown to protect various peptide regions (each shown in a different color) from hydrogen–deuterium exchange. Note that the resolution of the HDX experiments is at the peptide level, and reflects both fast and slow motions, up to the minute time-scale. (Right) Local Kullback–Leibler Divergence values between the apo and allosteric activator-bound ensembles are mapped onto the structure using PyMOL's "b-factor putty" preset. White indicates statistically insignificant divergence, and significant divergence values increase from blue to red. Most of the regions showing protection upon ligand binding also show statistically significant K–L divergence values.

4.2. Allosteric Inhibition by Lysine Acetylation in Mitochondrial 3-Hydroxy-3-methylglutaryl CoA Synthase 2.

Mitochondrial 3-hydroxy-3-methylglutaryl CoA synthase 2 (HMGCS2) is the rate-limiting enzyme in the synthesis of β-hydroxybutyrate and is normally acetylated at Lys310, Lys447, and Lys473, which inhibit its activity.[29] Both Lys447 and Lys473 are distant from the acetyl-CoA bound at the active site (Figure 2, gray spheres). Thus, the effects of these acetylations at Lys447 and Lys443 are allosteric in nature because they inhibit activity over a distance. For each construct, wild-type and mutant, we used five molecular simulations of 11–20 ns each (started with different random number seeds) on HMGCS2 in the deacetylated (activated) form and with acetylations at various lysine residues. These MD simulations showed that acetylation of specific lysines produced significant conformational and dynamical changes in HMGCS2.[30] As negative controls, two lysine residues whose acetylations do not inhibit activity were studied. In contrast to the other lysines, these did not show similar marked changes in structure and dynamics at the active site. The local Kullback–Leibler Divergence results showed a marked difference between the control lysine acetylations (acetylations at Lys306 or Lys437) that inhibited enzyme activity (acetylations at Lys447 and Lys473) (Figure 2). The control lysine acetylations did not show the pronounced divergences seen with the natural inhibitory lysine acetylations at positions 447 and 473.

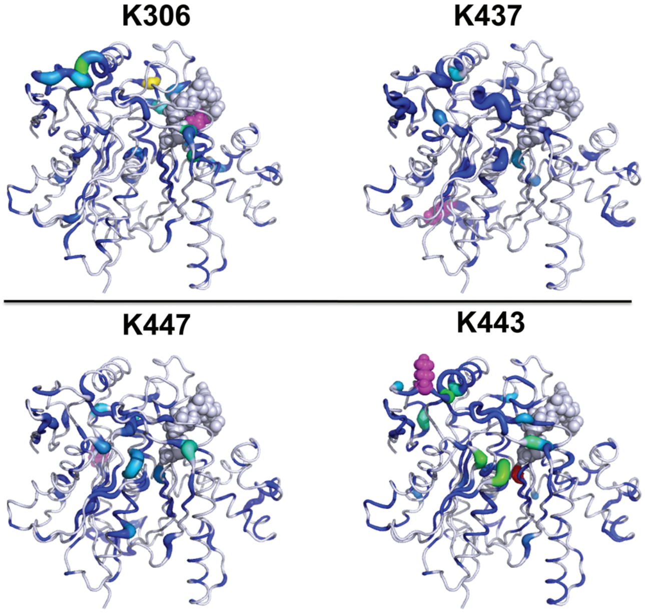

Figure 2. Kullback–Leibler Divergences show position-specific effects of lysine acetylation in HMGCS2. Local Kullback–Leibler Divergences between deacetylated and acetylated HMGCS2 conformational ensembles are given for different lysine acetylations (all are on the same scale). The lysine acetylated in each case is shown in purple spheres. (Top) Acetylation of Lys306 or Lys437 does not yield significant changes in structure and dynamics near the acetyl-CoA binding site (gray spheres) as assessed by the local Kullback–Leibler Divergence. (Bottom) In contrast to these negative controls, acetylation at lysine 447 or 443 causes substantial divergences proximal to the active site and the tail of the acetyl-CoA, and some background of divergences across the whole protein.

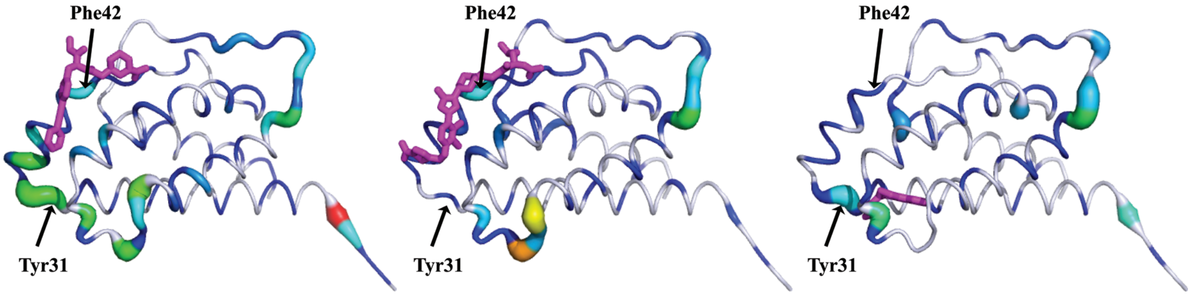

Figure 3. Kullback–Leibler Divergences between apo and ligand-bound IL-2 ensembles show differential allosteric effects. The local Kullback–Leibler Divergence between the apo IL-2 ensemble and various ligand-bound ensembles was calculated and mapped onto the apo structure. All panels are on the same scale, and ligands are superimposed for reference. These two binding sites were previously shown to be coupled through significant correlated torsional motions.[7] (Left) IL-2 with a micromolar ligand at the IL-2Rα site. (Center) IL-2 with an optimized nanomolar inhibitor at the IL-2Rα site featuring receptor-mimicking electrostatics.[31] (Right) IL-2 with an allosteric small-molecule fragment at a cryptic site.

Importantly, acetylation at Lys447 or Lys443 causes significant population shifts in the catalytic residues: in the loop containing the active site cysteine (Cys166), and in His301. Furthermore, these acetylations both cause substantial population shifts in a turn (239–241) near the acetyl-CoA tail, in Lys83 at the other end of the acetyl-CoA near the nucleotide ring, and in a helix-turn containing residues 380–385, which buttress the loop containing the active site cysteine (residues 163–168). In summary, the local Kullback–Leibler Divergence highlighted residues showing significant perturbations upon acetylation at Lys residues distant from the active site.

4.3. Communication between Small-Molecule Binding Sites in Interleukin-2. Interleukin-2 (IL-2) is a small cytokine that has been studied extensively as a model system for small molecules inhibiting protein–protein interactions. Binding of ligand to one site in IL-2 facilitates binding of a small-molecule fragment to a cryptic, transient pocket, which is gated by a loop on the opposite face from the four-helix bundle.[32] X-ray structures were unable to show how binding to one side affected binding at the other side, making it an interesting model system for studying small-molecule cooperativity; in prior work, we identified a putative allosteric network of residues coupling the binding sites using our MutInf method.[7] This putative allosteric network consisted of residues from the "bottom" of IL-2 in the orientation shown to the top (where IL-2Rα binds) along a "greasy core" consisting of Leu85, Phe78, Leu80, Tyr31, Met39, and Phe42, and a "polar network" consisting of Arg81, Gln74, Lys35, and Arg38 (which is then proximal to Met39 and Phe42 in the "greasy core"). We hypothesized that the Kullback–Leibler Divergence analysis would show significant population shifts in torsion angle distributions of residues implicated in the allosteric network by our previous mutual information analysis. We calculated the local, residue-by-residue Kullback–Leibler Divergence between apo (PDB: 1M47) and ligand-bound conformational ensembles from five 10 ns molecular dynamics simulations (Figure 3). Ligand-bound conformational ensembles analyzed here include a micromolar IL-2Rα-competitive inhibitor (PDB: 1M48), a nanomolar IL-2Rα-competitive inhibitor (PDB: 1PY2), and a weak fragment that only would bind in the presence of the micromolar inhibitor at the IL-2Rα-competitive site; cooperative binding of this fragment with the nanomolar inhibitor was not tested. The smaller inhibitor (left) but not the larger one (center) shows a substantial population shift on the helix-turn-helix at the fragment's binding site. Furthermore, the allosteric fragment (right) gives population shifts not only at its binding site but also in the helix containing hotspot residue Phe42 behind the IL-2Rα site ligands, as would be expected from thermodynamic linkage, indicating that the sampling was sufficient to observe an allosteric effect. As can be seen in Figure 3, population shifts are seen along this structurally contiguous network of residues upon binding of ligand at either site. In particular, Tyr31 seems to be an important mediator of allostery, as it is highlighted in both the left and right panels, whose respective ligands bind with positive cooperativity. Although this tyrosine does not directly contact the IL-2Rα-competitive inhibitor, the methionine in cyan located above it does contact the IL-2Rα-competitive inhibitor, and also contacts the allosteric fragment in the simulations. There are a number of polar residues proximal to Tyr31 that also show population shifts upon binding of allosteric fragment but that do contact the competitive inhibitor. Comparing these results to our previous study,[7] we find that all but two of the residues (all but Gln74 and Phe78) that were thought to be implicated in the putative allosteric network linking compound binding sites in our previous study showed statistically significant population shifts upon binding either an IL-2Rα-competitive inhibitor or an allosteric small-molecule fragment, and more specifically 5/6 "greasy core" residues and 3/5 "polar core" residues had Kullback–Leibler Divergence values in the top 25% of all residues' divergence values.

We also wondered whether our Kullback–Leibler Divergence values would highlight regions showing significant perturbations in experiments. In the case of IL-2, NMR chemical shift perturbations were available for the micromolar IL-2Rα-competitive inhibitor,[33] so we wondered whether regions of IL-2 showing chemical shift perturbations would also show significant Kullback–Leibler Divergence values. Such a comparison is complicated by the fact that the ligand itself, and especially its aromatic end that digs into a small pocket near a hotspot residue Phe42, will cause chemical shift perturbations in the protein residues apart from causing any shift in the protein dihedral distributions due to electronic effects and ring current effects involving aromatic residues. Nonetheless, we found in Figure 4 that most of the regions highlighted by the NMR chemical shift perturbations also showed significant Kullback–Leibler Divergences, with the notable exceptions being large Kullback–Leibler Divergence values in residues 4–5 and 97–102 where no significant NMR chemical shift perturbations were observed, and in residues 97–116 where the two signals do not overlap well. While we do not expect Kullback–Leibler Divergences to correlate with NMR chemical shift perturbations, it is interesting that these different and complementary measures seem to be picking up on regions

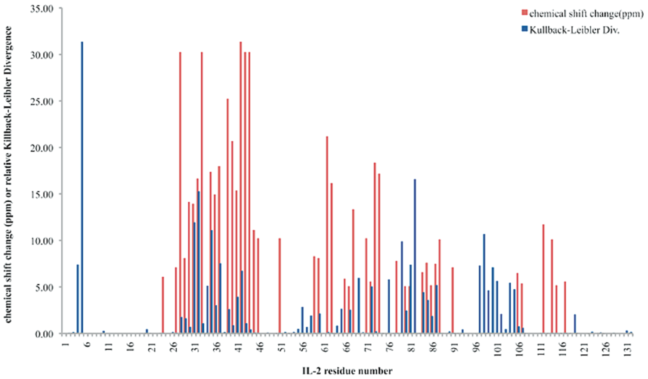

Figure 4. Kullback−Leibler Divergences highlight most regions showing NMR chemical shift perturbations in IL-2. Kullback−Leibler Divergences (blue), and NMR chemical shift perturbations over 5 ppm (red) are shown for each IL-2 residue for binding of the micromolar IL-2Rα-competitive inhibitor. Although quantitatively there is no significant residue-by-residue correlation between the magnitudes of these different measures of perturbation to the protein, we note that the two signals often highlight similar regions.

distant from the active site that show perturbation upon ligand binding.

4.4. pH Regulation of Talin.

Talin is an integrin-associated focal adhesion protein that binds actin with lower affinity at high pH and higher affinity at low pH. To investigate the mechanism by which pH change alters talin's structure, dynamics, and actin-binding ability, constant-pH molecular dynamics simulations of the I/LWEQ domain of talin1 without the C-terminal dimerization domain (PDB: 2JSW) were performed at pH 8.0 and pH 6.0 for 10 ns.[22] A histidine and nearby acidic residues with upshifted predicted pK_a values were hypothesized to constitute the pH sensor, and their protonation states were sampled during the constant-pH simulation. To test the importance of His2418, it was mutated to Phe. In NMR pH titrations, this mutant showed altered chemical shift perturbations, relative to wild type, and showed decreased F-actin binding in vitro and altered focal adhesion turnover in migrating cells. In this work, we apply the local Kullback−Leibler Divergence method to compare the conformational ensembles at these two different pH values for both wild-type and H2418F talin, using the pH 8.0 as the reference ensemble. We also wanted to compare the Kullback−Leibler Divergence and Jensen−Shannon Divergence values for this case, because our choice of reference state is somewhat arbitrary, and because the Jensen−Shannon Divergence is more robust to nonoverlap of the dihedral distributions, at a cost of being less sensitive due to the reference state being an average of the two ensembles. Because the H2418F mutant showed decreased F-actin binding at lower pH, we wondered whether this mutant would show less substantial population shifts at the actin binding site relative to wild-type talin. We found (Figure 5) that pH change caused substantial population shifts distant from the pH sensor in both cases, but that wild-type and H2418F talin in fact showed different patterns of population shifts in these actin binding site residues. The wild-type typically showed larger population shifts than the mutant in residues in the actin-binding site (boxed in red). The bottom of helices 1 and 3 in the wild-type shows substantial local Kullback−Leibler Divergences; in the NMR titration experiments,[22] both of these regions showed either chemical shift changes or line broadening.

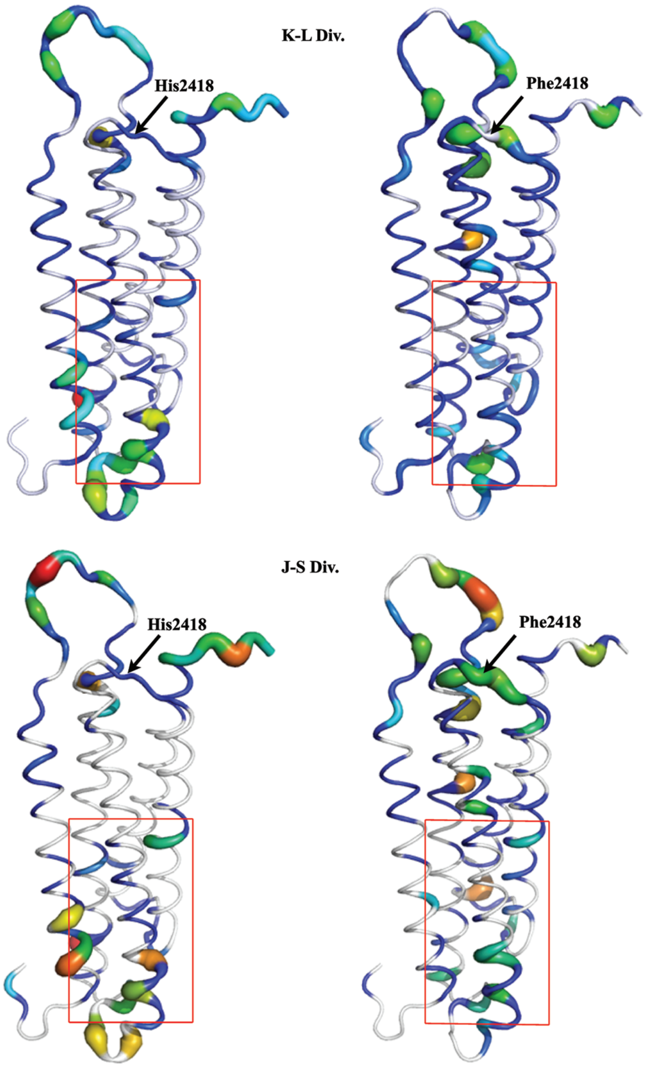

Figure 5. Wild-type and pH-sensor mutant talin show different population shifts upon pH change. Kullback−Leibler Divergences (top) and Jensen−Shannon Divergences (bottom) between the pH 8.0 ensemble and the pH 6.0 ensemble for talin are shown for wild-type (left) and H2148F talin (right). The actin-binding site region[22] is shown with a red box. These divergences highlight the region proximal to the pH sensor (at the top of the structure, with His2418 labeled) and the actin-binding site.

Given the population shifts in the putative pH sensor and actin binding site upon pH change in wild-type and H2418F talin, we wondered how protonation state changes in the pH sensor are propagated to the actin binding site. We suspect that a combination of correlated motions of charged residues and subtle rigid-body motions of the helices is responsible for coupling the pH sensor to the actin binding site. We observed subtle yet significant population shifts in the helices connecting the sites, which are qualitatively consistent with NMR chemical shift perturbations (Figure 6), that generally did not show large chemical shift perturbations in these residues, except in amides proximal to the pH sensor. Both the Kullback−Leibler Divergence and the Jensen−Shannon Divergence highlighted regions showing either chemical shift perturbations or line broadening in the titration, except proximal to the titrating His2418, where chemical shift perturbations would manifest direct electrostatic effects and not necessarily reflect substantial changes in the dihedral distributions.

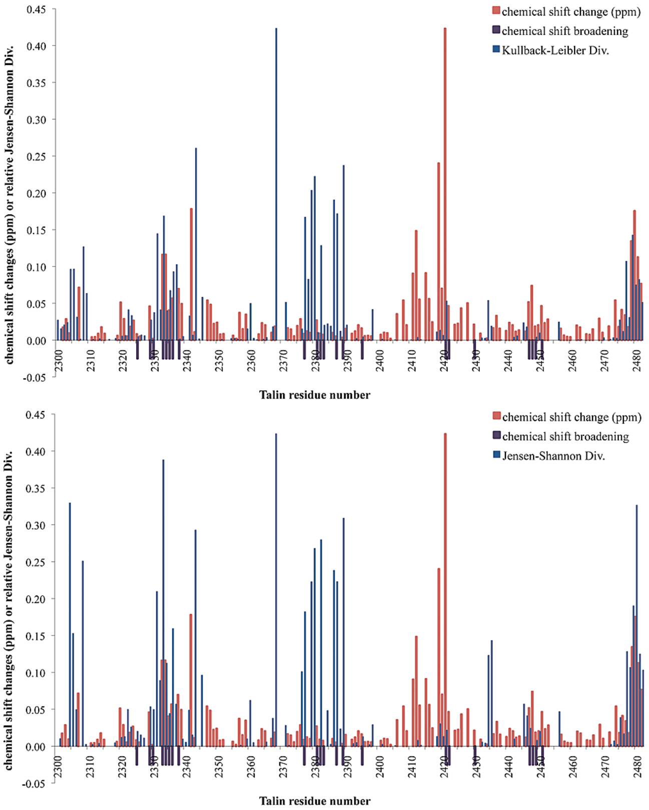

Figure 6. Kullback−Leibler Divergences and Jensen−Shannon Divergences between the pH 8.0 ensemble and the pH 6.0 ensemble highlight regions distant from the protonation sensor that show substantial NMR chemical shift perturbations or line shape broadening. (Top) Kullback−Leibler Divergence values and (bottom) Jensen−Shannon Divergence values for talin are compared to NMR chemical shift perturbations and line broadening in a pH titration. Amides whose normalized peak intensity ratio I_{pH8}/I_{pH6} were less than 51% are indicated with a purple bar below the x-axis. Chemical shift perturbations in ppm are given as abs($\delta\omega^{15}$N + 0.2$\delta\omega^{1}$H). Generally, both divergence measures highlight regions showing substantial chemical shift pertrubations or line broadening, with the notable exception of the region proximal to the titrating His2418.

5. DISCUSSION

We have developed a novel approach to comparing conformational ensembles that is grounded in thermodynamics, information theory, and statistics. We use the Kullback−Leibler Divergence to quantify changes in torsion angle probability distributions, which reflect biologically relevant processes such as side chain rotamer flips, changes in local secondary structure, etc. Inspired by previous work, we developed the Kullback−Leibler Divergence expansion, which provides an approximation the Kullback−Leibler Divergence of whole molecules (proteins in this work) in terms of marginal probability density functions involving far fewer degrees of freedom. In this work, we have found that even the first-order terms can give considerable qualitative insight into which residues are most affected by perturbations such as ligand binding or pH change (i.e., proton binding). The expansion presented here to approximate the Kullback−Leibler Divergence for a macromolecule can include couplings beyond the pairwise level. In contrast, an exact expression for the Kullback−Leibler Divergence at second order for a harmonic system was previously presented by Ming and Wall.[10] While Ming and Wall's exact second-order method was based on a normal mode model, the present approach leverages data from molecular dynamics simulations that sample the free energy landscape and then account for anharmonicities in the analysis through a distribution-free measure of changes to probability densities. While the method presented here is approximate rather than exact, the use of molecular dynamics simulations provides more realistic dynamics that might reveal linkages between sites separated by long distances and through mechanisms other than vibrational couplings through semirigid elements, which normal mode models can often detect. In the initial applications considered here, we see evidence for allosteric communication propagating through helices, acting as semirigid elements. In these cases, the conformational ensembles of residues on opposite ends of the helix (or sheet) are perturbed by the same ligand or mutation, but the residues in the semirigid element can be minimally perturbed. We also observe many significant divergences in surface polar residues; this suggests that these residues may play a role in coupling binding at one site to a change in structure and/or dynamics at another site. As these surface polar side-chains are often not part of evolutionarily conserved networks,[34] their ability to propagate these kinds of perturbations may lie in the sum effects of multiple residues working in a parallel fashion. We speculate that, in such cases, the specific amino acid identity is less important than their physical properties, such as holding a charge or strong dipole that can reorient with the help of a flexible linker. A similar role for correlated protein side chain motions in mediating long-range couplings was suggested by DuBay and Geissler.[35] There are several algorithmic improvements that could be made to our approach. Multiple calculations at different histogram bin sizes could be used, and an optimal histogram size chosen for each degree of freedom; such an approach has been shown to lead to more accurate entropy calculations.[36] A k-nearest neighbor (KNN) approach could also be used to calculate the Kullback−Leibler Divergence.[37] Our residue-level analysis of the local Kullback−Leibler Divergence could be augmented by including second-order terms (i.e., the mutual divergence) within residues, which could benefit from adaptive partitioning, as in our previous work on mutual information.[7]

■ APPENDIX

Robust Histogram Estimate of Kullback−Leibler Divergence Using Renyi Generalized Divergence

To obtain a finite-sample size correction to the Kullback−Leibler Divergence, we adapt the derivation presented by Grassberger.[38] Although this was not used in the applications shown here, it is provided as an option in the program and is provided here for completeness and for possible inclusion in other code packages. We consider the Kullback−Leibler Divergence as a limit of the Renyi Generalized Divergence:

$$\mathrm{KL}_i = \lim_{\alpha \to 1} D_\alpha(P \| P^*) \tag{29}$$

where

$$D_\alpha(P \| P^*) = \frac{1}{(\alpha - 1)} \ln\left(\sum_{i=1}^{n} \frac{p_i^\alpha}{(p_i^*)^{\alpha - 1}} \right) \tag{30}$$

For finite sample sizes, there will be some uncertainty in the p_i. Considering the actual histogram counts, we write:

$$p_i^\alpha = \left(\frac{\langle n_i \rangle}{N}\right)^\alpha, \; (p_i^*)^\alpha = \left(\frac{\langle n_i^* \rangle}{N}\right)^\alpha \tag{31}$$

To obtain $\langle n \rangle^\alpha$, we assume a Poisson distribution for n_i in successive realizations (i.e., assuming we are using a fine enough discretization such that $p_i \ll 1$). For a positive integer α, we would then have

$$\langle n \rangle^\alpha = \left\langle \frac{n!}{(n-\alpha)!} \right\rangle \tag{32}$$

However, in the limit as α approaches 1, we need a continuous analogue using Γ functions. Grassberger found an asymptotic expansion for $\langle n_i \rangle^\alpha$ and showed that two terms gave numerically robust results for Shannon entropies.

$$\langle n \rangle^\alpha = \frac{\Gamma(n+1)}{\Gamma(n-a+1)} - \frac{(-1)^n \Gamma(a+1)\sin(\pi a)}{\pi(n+1)} \tag{33}$$

This same approximation is used in our previously published MutInf method.[7] Next, we use this expression for $\langle n \rangle$ and evaluate the Renyi Generalized Divergence in the $\alpha \to 1$ limit to give us the Kullback–Leibler Divergence. Invoking L'Hopital's Rule, we obtain:

$$\lim_{\alpha \to 1} D_\alpha(P\|Q) = \lim_{\alpha \to 1}\left(\sum_{i=1}^{n\text{bins}} \frac{Nf(n_i^*, \alpha - 1)}{f(n_i, \alpha)} \frac{\partial}{\partial \alpha} \sum_{i=1}^{n\text{bins}} \frac{f(n_i, \alpha)}{Nf(n_i^*, \alpha - 1)}\right) \tag{34}$$

$$\lim_{\alpha \to 1} D_\alpha(P\|Q) = \sum_i^{n\text{bins}} \frac{\Psi(n_i)n_i n_i^* + (-1)^{n_i} n_i^* + (-1)^{n_i^*} n_i n_i^* - \Psi(n_i^*)n_i n_i^* - n_i}{n_i^*} \tag{35}$$

However, this expression is not numerically robust in practice, so we truncate the expression for $\langle n \rangle^\alpha$ at the first term:

$$\langle n \rangle^\alpha = \frac{\Gamma(n+1)}{\Gamma(n-a+1)} \tag{36}$$

which then provides a more robust estimate for $D_\alpha(P\|Q)$:

$$\text{KL}_1 = \lim_{\alpha \to 1} D_\alpha(P\|Q) = \sum_i^{n\text{bins}} \frac{n_i}{N}\left(\Psi(n_i) - \Psi(n_i^*) - \frac{1}{n_i^*}\right) \tag{37}$$

Using a series approximation of the digamma function, $\Psi(x) \approx \ln(x) - (1/2x)$, it can be readily seen that the Kullback–Leibler Divergence in eq 23 is recovered along with a correction term that decreases in size as histogram counts increase.

AUTHOR INFORMATION

Corresponding Author

*E-mail: clmcclendon@ucsd.edu.

Present Addresses

[||]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA 92093, USA.
[⊥]Pfizer Global Research and Development, Groton, CT 06340, USA.

Notes

The authors declare the following competing financial interest(s): Matthew P. Jacobson is a consultant for Schrodinger, LLC and for Pfizer, Inc.

ACKNOWLEDGMENTS

We would like to thank Ken Dill for inspiring this work, and Steve Presse, Michael Gilson, and Andrew Fenley for helpful advice and comments on the manuscript. This work was supported by NSF Teragrid Allocation TG-MCB090109 and the Texas Advanced Computing Center (TACC). C.L.M. was supported by a PhRMA Foundation Informatics Fellowship and is currently supported by NIH grant F32 GM099197-01. This project was supported by the HARC center, NIH grant P50-GM082250. In the course of this work, analyses were performed with PyMOL (Schödinger, LLC) and the UCSF Chimera package.[39] Chimera is developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from the National Institutes of Health (National Center for Research Resources grant 2P41RR001081, National Institute of General Medical Sciences grant 9P41GM103311). M.P.J. is a consultant to Schrodinger, LLC and Pfizer, Inc.

REFERENCES

(1) Morcos, F.; Chatterjee, S.; McClendon, C. L.; Brenner, P. R.; López-Rendón, R.; Zintsmaster, J.; Ercsey-Ravasz, M.; Sweet, C. R.; Jacobson, M. P.; Peng, J. W.; Izaguirre, J. A. *PLoS Comput. Biol.* **2010**, *6*, e1001015.

(2) Ramanathan, A.; Savol, A. J.; Langmead, C. J.; Agarwal, P. K.; Chennubhotla, C. S. *PLoS One* **2010**, *6*, e15827.

(3) Koyama, Y. M.; Kobayashi, T. J.; Tomoda, S.; Ueda, H. R. *Phys. Rev. E* **2008**, *78*, 046702.

(4) Lange, O. F.; Grubmüller, H. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 1294–1312.

(5) Bradley, M. J.; Chivers, P. T.; Baker, N. A. *J. Mol. Biol.* **2008**, *378*, 1155–1173.

(6) Lange, O. F.; Grubmüller, H. *Proteins: Struct., Funct., Bioinf.* **2006**, *62*, 1053–1061.

(7) McClendon, C. L.; Friedland, G.; Mobley, D. L.; Amirkhani, H.; Jacobson, M. P. *J. Chem. Theory Comput.* **2009**, *5*, 2486–2502.

(8) Qian, H. *Phys. Rev. E* **2001**, *63*, 042103.

(9) Wall, M. E. *AIP Conf. Proc.* **2006**, *851*, 16–33.

(10) Ming, D.; Wall, M. E. *Proteins: Struct., Funct., Bioinf.* **2005**, *59*, 697–707.

(11) Ming, D.; Cohn, J.; Wall, M. *BMC Struct. Biol.* **2008**, *8*, 5.

(12) Matsuda, H. *Phys. Rev. E* **2000**, *62*, 3096.

(13) Killian, B. J.; Kravitz, J. Y.; Gilson, M. K. *J. Chem. Phys.* **2007**, *127*, 024107–16.

(14) Somani, S.; Killian, B. J.; Gilson, M. K. *J. Chem. Phys.* **2009**, *130*, 134102.

(15) Killian, B. J.; Kravitz, J. Y.; Somani, S.; Dasgupta, P.; Pang, Y.-P.; Gilson, M. K. *J. Mol. Biol.* **2009**, *389*, 315–335.

(16) Karchin, R.; Kelly, L.; Sali, A. *Pac. Symp. Biocomput.* **2005**, 397–408.

(17) Sakuraba, S.; Joti, Y.; Kitao, A. *J. Chem. Phys.* **2010**, *133*, 185102.

(18) Sweet, , C. R.;; Petrone, , P.; ; Pande, , V. S.;; Izaguirre, , J. A. *J. Chem. Phys.* **2008**, *128*, 145101.

(19) Izaguirre, J. A.; Sweet, C. R.; Pande, V. S. *Pac. Symp. Biocomput.* **2010**, 240–251.

(20) King, B. M.; Tidor, B. *Bioinformatics* **2009**, *25*, 1165–1172.

(21) Crooks, G. E. *Phys. Rev. Lett.* **2007**, *99*, 100602.

(22) Srivastava, J.; Barreiro, G.; Groscurth, S.; Gingras, A. R.; Goult, B. T.; Critchley, D. R.; Kelly, M. J. S.; Jacobson, M. P.; Barber, D. L. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 14436–14441.

(23) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. *Intermolecular Forces*; D. Reidel Publishing Co.: Dordrecht, 1981; pp 331–342.

(24) Bowers, K. J.; Chow, E.; Huafeng, X.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Yibing, S.; Shaw, D. E. *Proc. ACM/IEEE Conf. Supercomputing (SC06)* **2006**, 43.

(25) Berendsen, H. J. C.; Postma, J. P. M; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(26) Martyna, G. J.; Tobias, D. J.; Klein, M. L. *J. Chem. Phys.* **1994**, *101*, 4177−4189.
(27) Tuckerman, M.; Berne, B. J.; Martyna, G. J. *J. Chem. Phys.* **1992**, *97*, 1990−2001.
(28) Hindie, V.; Stroba, A.; Zhang, H.; López-Garcia, L. A.; Idrissova, L.; Zeuzem, S.; Hirschberg, D.; Schaeffer, F.; Jørgensen, T. J. D.; Engel, M.; Alzari, P. M.; Biondi, R. M. *Nat. Chem. Biol.* **2009**, *5*, 758−764.
(29) McGarry, J. D.; Foster, D. W. *Annu. Rev. Biochem.* **1980**, *49*, 395−420.
(30) Shimazu, T.; Hirschey, M. D.; Hua, L.; Dittenhafer-Reed, K. E.; Schwer, B.; Lombard, D. B.; Li, Y.; Bunkenborg, J.; Alt, F. W.; Denu, J. M.; Jacobson, M. P.; Verdin, E. *Cell Metab.* **2010**, *12*, 654−661.
(31) Thanos, C. D.; DeLano, W. L.; Wells, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 15422−15427.
(32) Hyde, J.; Braisted, A. C.; Randal, M.; Arkin, M. R. *Biochemistry* **2003**, *42*, 6475−83.
(33) Emerson, S. D.; Palermo, R.; Liu, C.-M.; Tilley, J. W.; Chen, L.; Danho, W.; Madison, V. S.; Greeley, D. N.; Ju, G.; Fry, D. C. *Protein Sci.* **2003**, *12*, 811−822.
(34) Halabi, N.; Rivoire, O.; Leibler, S.; Ranganathan, R. *Cell* **2009**, *138*, 774−86.
(35) DuBay, K. H.; Geissler, P. L. *J. Mol. Biol.* **2009**, *391*, 484−497.
(36) Baron, R.; Hünenberger, P. H.; McCammon, J. A. *J. Chem. Theory Comput.* **2009**, *5*, 3150−3160.
(37) Piro, P.; Anthoine, S.; Debreuve, E.; Barlaud, M. *International Workshop on Content-Based Multimedia Indexing*, 2008; pp 230−235.
(38) Grassberger, P. *Phys. Lett. A* **1988**, *128*, 369−373.
(39) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. *J. Comput. Chem.* **2004**, *25* (13), 1605−1612.