THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

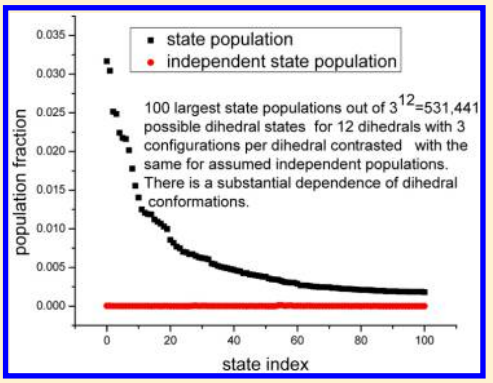
Dihedral Angle Entropy Measures for Intrinsically Disordered Proteins

Robert I. Cukier*

Department of Chemistry, Michigan State University, East Lansing, Michigan 48824-1322, United States

Supporting Information

ABSTRACT: Protein stability is based on a delicate balance between energetic and entropic factors. Intrinsically disordered proteins (IDPs) interacting with a folded partner protein in the act of binding can order the IDP to form the correct functional interface by decrease in the overall free energy. In this work, we evaluate the part of the entropic cost of ordering an IDP arising from their dihedral states. The IDP studied is a leucine zipper dimer that we simulate with molecular dynamics and find that it does show disorder in six phi and psi dihedral angles of the N terminal sequence of one monomer. Essential to ascertain is the degree of disorder in the IDP, and we do so by considering the entire, discretized probability distribution function of N dihedrals with M conformers per dihedral. A compositional clustering method is introduced, whereby the $N_S = N^M$ states are formed from the Cartesian product of each dihedral's conformational space. Clustering is carried out with a version of a k-means algorithm that accounts for the circular nature of dihedral angles. For the 12 dihedrals each found to have three conformers, among the resulting 531441 states, their populations show that the first 100 (500) most populated states account for ∼65% (∼90%) of the entire population, indicating that there are strong dependencies among the dihedrals' conformations. These state populations are used to evaluate a Kullback−Leibler divergence entropy measure and obtain the dihedral configurational entropy S. At 300 K, $TS \sim 3$ kcal/mol, showing that IDP entropy, while roughly half that would be expected from independently distributed dihedrals, can be a decisive contributor to the free energy of this IDP binding and ordering.



1. INTRODUCTION

The stability of a folded protein, on the order of 5−15 kcal/mol, arises from a subtle compromise among disparate forces.[1−3] This rather small stability most likely is a consequence of the requirements of spatially and temporally organized folding/unfolding events in cellular function. While folded proteins are essential for catalytic activity, the role of disordered states in proteins has become manifest more recently. Intrinsically disordered proteins (IDPs) that may sample a large conformational space by rapid interconversions among a large number of states are known to be essential for cell function.[4−7] The crucial roles of IDPs in the regulation of transcription and translation for the purpose of cell signaling are now appreciated.[4−8] For protein−protein interactions, where one protein is initially ordered and the other disordered, the act of binding may order the partner protein to form the correct functional interface by decrease in the interfacial free energy. Whether this takes place via "induced fit" whereby the interaction of the disordered with the ordered partner induces a restructuring or by "conformational selection", whereby a small, ordered population is selected to bind, or something in between, is always an issue.[6,9] Also key is the timing and ordering of binding and folding events.[8,10−12]

One class of IDP scenarios arises in protein−protein interactions, where the disorder is in a loop region. Another kind of IDP scenario arises for leucine zippers that are composed of dimerized monomers.[13] The formation of the dimer interface is thought to proceed via a trigger sequence,[14,15] whereby the ordered C termini of the monomers interact while the N terminus of one or both monomers is separated from its partner and is disordered to a certain extent until it does fully interact with its partner and becomes ordered. Leucine zipper folding and stability has been experimentally probed by calorimetry, circular dichroism, hydrogen-exchange kinetics, and NMR.[15−21] Leucine zippers are an integral part of bZIP (basic region, leucine zipper) DNA binding proteins and are important to transcriptional regulation in eukaryotes.[13,22−24] In this work, we analyze the amount of disorder in an N terminal sequence of a leucine zipper, as obtained from a molecular dynamics (MD) simulation, and assess its contribution to the entropy relative to when it is bound in its dimerized form.

From a free-energy perspective, the role of an IDP and its potential for ordering when interacting with another protein

Received: October 10, 2014
Revised: February 13, 2015
Published: February 13, 2015

DOI: 10.1021/jp5102412
J. Phys. Chem. B 2015, 119, 3621−3634

or general binding partner will have a strong entropic component. Essential to all these considerations is the degree of disorder in the IDP.[7] As an IDP, it is useful to have some but not too much disorder to do the fine-tuning required for the equilibrium between, for example, dimerized and dissociated states, particularly in response to various signals required for the timing of events such as translocation. What is needed, then, is some measure of disorder/order in IDPs. That requires a way to define states in an IDP and, as well-known, dihedral angles are an appropriate coordinate set for peptides and for conformations of loops in proteins.[25−30] When a dihedral changes from one stable basin to another, it can induce a large, global conformational change. Dihedrals are also useful in this regard because their sampling tends to be confined to relatively distinct regions of configuration space as revealed by, for example, Ramachandran plots of backbone phi and psi dihedrals.[31]

Then, the issue becomes which dihedral states are sampled and how to count them, with the difficulty that there are many dihedrals so the state space is intrinsically high-dimensional. Furthermore, it is most likely that the dihedrals in an IDP will have strong dependences, the configurational probability distribution of a particular dihedral will depend on the configurations of the other dihedrals. Then the associated entropy is very different than that of a "random coil" where, if every backbone dihedral were independent, the N dihedral probability density function (pdf) $P(\mathbf{x}) \equiv P(x_1,x_2,..,x_N)$ is the product of each dihedral's pdf: $P(x_1, x_2, ..., x_N) = \prod_{n=1}^{N} P(x_n)$. Dependence will reduce entropy and will reduce the free-energy cost for the IDP-to-structured (e.g., helix) transition. Other things being equal, increasing amounts of dependence will reduce the cost of forming the structured state. Thus, one critical task is to establish the degree of dependence among the dihedral states. That there can be strong dependencies among the dihedral angles contributing to a state may be a consequence of the feature that several dihedrals can undergo compensating conformational changes and lead to small changes in the overall conformation.[32−34] In principle, a small overall conformational change, which does imply a correlation among a set of dihedrals, may be less costly in terms of free energy than conformational changes with a large configuration space displacement.

Our focus will be on configurational contributions to entropy arising from dihedral angles. There are of course other, significant contributions to the entropy. The distinction between "hard" (bond and bond angle) and "soft" (dihedral) coordinates was made a long time ago,[25−27] and their respective contributions to entropy may be considered as additive.[35] For the hard coordinates, various quasi-harmonic entropy expressions that assume multivariate Gaussian fluctuations around some conformational basin have been given.[36−40] For dihedral coordinates that involve intrinsically nonharmonic transitions among multibasin conformations, a normal coordinate decomposition is not appropriate. If the coordinates are assumed independent then the entropy is additive over each coordinate.[25] A number of methods that go beyond independence have been presented. Kirkwood's closure of the BBGKY hierarchy via the superposition approximation can be used as a systematic method to express higher-order correlation functions in terms of lower-order ones.[41] Cluster decompositions of the mutual information entropy have been obtained to estimate the influence of dependencies. The Kullback−Leibler divergence[42] has been expressed in term of a generalized Kirkwood superposition approximation and used to compare protein and protein ligand-bound conformational ensembles.[43] The maximum information spanning tree (MIST) method[44] was used to investigate ligand binding entropies,[45] protein side-chain entropies, and their connection with NMR order parameters,[46,47] protein entropy−enthalpy transduction,[48] and the effect of correlations on entropy in binding and conformational changes in proteins.[49] A mutual information expansion (MIE)[50] was developed and applied to the role of correlations in small molecule binding to a protein[51] and protein−ligand binding.[52] Another approach to the role of correlation, the multibody local approximation (MLA), was developed and applied to a peptide.[53] It has been integrated with the CENCALC program[54] and used to assess convergence of MD trajectory data.[55]

A desirable feature of a scheme to define an appropriate entropy measure is to maximize the "contrast" of the states that it uses. As noted above, dihedrals are good descriptors of conformations of peptides and of loops in otherwise stable proteins. When a dihedral changes from one stable basin to another, it can induce a large, global conformational change. Given an appropriate dissimilarity measure, states can be defined with the use of a clustering algorithm.[56] For proteins, clustering is typically done in the collective space of a set of degrees of freedom.[28,57−60] For example, the root-mean-square deviation (RMSD) is the classic one that sums over the squared distances between snapshots of a set of atom coordinates and those of some reference structure. We refer to this collective space clustering as global clustering. It is also often done for dihedrals.[57,58] Here, we introduce an alternative procedure and, in this work, apply it to dihedral angles. Namely, for dihedrals, define a state NDH = $1DH_1 \times 1DH_2 \times \cdots \times 1DH_N$, where $1DH_n$ spans the nth dihedral's (n = 1, 2, ..., N) conformations and NDH is their Cartesian direct product. We will refer to this state definition as compositional clustering. Each $1DH_n$ is composed of at least two conformers. In this method, the trajectory of each dihedral is individually clustered into a set of conformers. The NDH states are then defined as all possible n-fold products (the composition) of these one-dihedral clusters. Compositional clustering is a very different state definition than used for RMSD clustering, where the states are defined collectively. (The Cartesian coordinate analog of compositional clustering would be to cluster, for example, a set of distances and take the composition of them to form a state).

If the focus is on the N-dimensional dihedral pdf, there is the problem that if a continuous space is used (with binning, of course), for example, 36 bins for each dihedral on (−180°, +180°) then there are 36^{NDH} possible states. Clearly this becomes impractical for more than a few dihedrals. Not only does the dimensionality explode but also the amount of data required for adequate statistics becomes too large. Because dihedral probability distributions tend to be localized into two or three peaks, reflecting their specification as rotamers, a discretization based on this feature can be used to construct configurational entropy from the composition of the N discretized dihedrals. Thus, we will discretize the state space by defining states as strings formed from a composition of the discretized dihedrals. For example, for three dihedrals each with two possible conformers, as typical of backbone phi and psi dihedrals, there are $2^3 = 8$ states. For computational manipulation, it is useful to define these states as strings. That

DOI: 10.1021/jp5102412
J. Phys. Chem. B 2015, 119, 3621−3634

is, for the $2^3 = 8$ states define strings $(I_1I_2I_3)$, where $I_i \in \{0,1\}$ $(i = 1,2,3)$. Thus, we will use a discretized state space, which is well-suited to dihedrals that is the mainstay in well-defined regions of angle space. Throughout this work, "states" will denote the many-dihedral combinations formed from each dihedral's possible "conformations".

To carry out this discretization, a clustering algorithm is required. A popular one is *k*-means clustering.[56] However, in a high-dimensional space, *k*-means has the drawback that it becomes so compute-intensive that typically,[56] and certainly with MD trajectory data,[57] only a subset of the trajectory snapshots will be used for the clustering. Clearly, with our composition version, that is not an issue. Each dihedral is clustered separately, and this is an efficient operation.

Clustering a dihedral angle introduces an issue particular to circular data on (−180°, +180°) with its periodicity versus Cartesian space distances. Computing a mean angle and distance (dissimilarity) of a dihedral snapshot to the mean angle, required for centroid-based, or other, clustering methods requires a distance definition based on circular statistics. We will use the one introduced by Fisher as it provides a correct definition of means for circular data.[61] There can be a difference between the number of clusters that would result if an ad hoc procedure based on the dihedral angular histograms constructed from trajectory snapshots were used versus circular statistics-based clustering. These differences can lead to different state spaces and are instructive to explore. We will use *k*-means clustering based on the Fisher distance definition and refer to the method as *k*-means Fisher clustering.

Because our focus is on the *N*-dimensional pdf and the dependence among the dihedrals, it is natural to use a relative entropy measure, the Kullback−Leibler divergence (KLD)[42] to quantitate the extent of dependence. The KLD measure has been used in a number of protein contexts to compare conformational distributions.[28,62−64] When the KLD is specialized to one distribution being the independent distribution, the KLD is the mutual information.[42] In contrast, with a Pearson correlation coefficient that indicates a degree of correlation, the mutual information provides an amount of dependence. On the basis of the discrete states and their populations that are generated from an analysis of the leucine zipper simulation trajectory data, it is straightforward to generate KLD values. They provide an indication of how far from independent the states are. The dihedral configurational entropy values obtained here provide contributions to the free energy that are on the same scale as protein-binding free energies and thus can be instrumental in IDP ordering.

The remainder of this paper is organized as follows. Section 2 presents the MD protocol to generate a leucine zipper trajectory where the N terminal sequence becomes disordered in its phi and psi backbone dihedrals. The compositional *k*-means Fisher clustering algorithm is detailed and the KLD entropy measure defined. Our workflow is outlined, and a computational algorithm to construct the dihedral state populations is described. In Section 3, we first introduce synthetic data that clarifies the distinction between ad hoc clustering in the dihedral angle space and Fisher-based clustering. Then, the dihedral simulation data is presented, the state space and their probabilities found, and the KLD evaluated. To aid in the interpretation of the KLD values, three evocative examples are constructed. Section 4 presents a discussion of our results, and Section 5 provides some concluding remarks.

2. METHODS

2.1. MD Protocol. The CUKMODY protein molecular dynamics code that uses the GROMOS96[65] force field was used for all simulations; the conditions are detailed elsewhere.[66] The starting GCN4-p1 leucine zipper dimer configuration was obtained from its X-ray structure (PDB accession code 2ZTA).[67] The simulations were carried out in a box with 59.1851 Å sides with SPC waters. Configurational sampling was enhanced with a Hamiltonian Temperature Replica Exchange Method (HTREM) that scales the Hamiltonian in both potential and kinetic energies.[66] The potential scaling is carried out only for the protein−protein and protein−solvent interactions and the kinetic scaling only for the protein degrees of freedom. By limiting the number of scaled degrees of freedom, a smaller number of systems can be used relative to temperature REM, where all degrees of freedom are scaled. For the current purposes, the scaling of the replicas was kept quite close to unity, in effect generating a set of independent trajectories, all at essentially the same effective temperature, much like initiating independent trajectories using different initial velocity distributions.

Two simulation trajectories were generated. The first corresponds to a simulation of the dimer at its "normal" dimer interface separation, whereby the structure is a fluctuating version of the crystal structure. Using eight replicas, eight independent trajectories were generated. Each trajectory was run for 4 ns and in total provided 32000 snapshot sampled every picosecond. This provided a baseline data set for the dihedral conformations of the monomers. In particular, the N terminal residues of monomer 1, residues 3−15, along with the remaining residues 16−31 (as well as monomer 2) are largely alpha helical. (The first two residues are excluded because they extensively fluctuate even in the well-bound dimer.) The second simulation incorporated monomer−monomer restraints on the N terminal sequence part of the leucine zipper. By pushing this part of the dimer apart using these restraints, it was found that the N terminus of one of the monomers, monomer 1, "melted". That is, for residues 3−8 corresponding to roughly one-half of the N terminal sequence, the alpha helical character found for the first simulation is disrupted, and this provided the data for the multiconformer dihedral sampling. The pushing apart simulation again used eight replicas of 4 ns each. Subsequently, two 4 ns runs using eight replicas, separated by a 2 ns run, were carried out in order to compare their results for evaluation of the statistical quality of the data. Trajectories of the phi and psi dihedral angles were generated using ANALYZER.[68]

2.2. Dihedral *k*-Means Fisher Clustering Algorithm. NDH, the *N*-dimensional dihedral discrete state space that will be used here is formed from the Cartesian product of one-dimensional dihedral spaces

$$\mathrm{NDH} = 1\mathrm{DH}_1 \times 1\mathrm{DH}_2 \times \ldots \times 1\mathrm{DH}_N \quad (1)$$

with $1\mathrm{DH}_n = \{1, 2, \ldots, M_n\}$ with M_n a finite integer. (Naturally, the dihedral multiplicity will most likely satisfy $1 \leq M_n \leq 3$.) The states in NDH denoted by $s = (1, 2, \ldots, N_S)$ number

DOI: 10.1021/jp5102412
J. Phys. Chem. B 2015, 119, 3621−3634

$$N_S = \prod_{n=1}^{N} M_n \tag{2}$$

growing exponentially with the number of dihedrals. Use of the Cartesian product definition of the state space then permits construction of the NDH states by a one-dimensional k-means clustering algorithm that will clearly be much more efficient than an N-dimensional clustering algorithm.

To carry out clustering of each dihedral angle, it is imperative to use an appropriate definition of a mean direction, such as the one introduced by Fisher.[61] Directions in "Fisher" space are defined as follows. For a given dihedral angle from the ith of n trajectory snapshots, set

$$X_i = \cos(\theta_i), \quad Y_i = \sin(\theta_i) \tag{3}$$

and define

$$X = \sum_{i=1}^{n} X_i, \quad Y = \sum_{i=1}^{n} Y_i \tag{4}$$

and

$$\cos\bar{\theta} = X/(X^2 + Y^2)$$
$$\sin\bar{\theta} = Y/(X^2 + Y^2) \tag{5}$$

where

$$\bar{\theta} = \begin{cases} \tan^{-1}(Y/X) Y > 0, & X > 0 \\ \tan^{-1}(Y/X) + \pi X < 0 & \\ \tan^{-1}(Y/X) + 2\pi Y < 0, & X > 0 \end{cases} \tag{6}$$

is the principal value of the arctangent. A geometric construction shows that $\bar{\theta}$ is the resultant direction obtained by adding each sample angle vectorially in $\mathbf{R}^2$.

For application to a k-means algorithm,[56] define the kth centroid $(k = 1, 2, ..., K)$ as

$$\mathrm{mX}^{it}[k] = \sum_{X_i \in S_k^{it}} X_i / \sum_{X_i \in S_k^{it}} 1$$
$$\mathrm{mY}^{it}[k] = \sum_{Y_i \in S_k^{it}} Y_i / \sum_{Y_i \in S_k^{it}} 1 \tag{7}$$

where $S^{it}[k]$ denotes the current cluster assignment of each (X_i,Y_i):

$$(X_i, Y_i) \in S^{it}[k] \tag{8}$$

Initially, the k centroids must be assigned to start off the iteration. On the it'th iteration, each θ_i is assigned to its current cluster $S^{it}[k]$ by assigning each (X_i,Y_i) to its closest current centroid using the Cartesian distances $d_i[k]$ defined as

$$d_i[k] \equiv [(X_i - \mathrm{mX}^{it}[k])^2 + (Y_i - \mathrm{mY}^{it}[k])^2]^{1/2}, (k = 1, 2, ..., K) \tag{9}$$

between snapshots and current centroids. Thus,

$$S_i^{it}[k] = \{(X_i, Y_i) | d_i[k] \le d_i[k'], \quad (k, k' = 1, 2, ..., K)\} \tag{10}$$

The centroids are re-evaluated using eq 7, and the scheme terminates when

$$\mathrm{m}^{it+1}[k] = \tan^{-1}(\mathrm{mY}^{it+1}[k], \mathrm{mX}^{it+1}[k]) \tag{11}$$

satisfies

$$\mathrm{abs}(\mathrm{m}^{it+1}[k]) - \mathrm{abs}(\mathrm{m}^{it}[k]) < \mathrm{tol}, \quad (k = 1, 2, ..., K) \tag{12}$$

to some tolerance, tol.

Two aspects of k-means are worth noting. First, whatever the metric used, there can be a sensitivity to the initial choice of the centroids. Second, specific to the nonlinear map between angle and the construction of the centroids based on Cartesian space distances, what could appear as an unresolved set of data when histogrammed in angular space may be resolved in the k-means Fisher clustering.

2.3. KLD Entropy Measure. The relative entropy or Kullback–Leibler divergence (KLD)[42,69] between two pdfs, $p(\mathbf{x})$ and $q(\mathbf{x})$, the latter a reference pdf, for an N degrees of freedom vector of random variables $\mathbf{x} = (x_1, x_2, ..., x_N)$ whose possible values are denoted by $\mathbf{X} = (X_1,X_2, ..., X_N)$ is

$$\mathbf{X} = (X_1, X_2, ...X_N) \text{ is } D(p\|q) = \sum_{\mathbf{x}\in\mathbf{X}} p(\mathbf{x}) \log\frac{p(\mathbf{x})}{q(\mathbf{x})} (D(p\|q) \ge 0) \tag{13}$$

For our purposes, the random variables x_i are the dihedral angles and will be assumed to take on a set of a small number of discrete values (obtained by the clustering algorithm) denoted by X_i. The reference of our interest is the independent pdf; $q(\mathbf{x}) \equiv p^{\mathrm{ind}}(\mathbf{x}) = \prod_{n=1}^{N} P(x_n)$

so that

$$D(p\,\|\,p^{\mathrm{ind}}) = \sum_{\mathbf{x}\in\mathbf{X}} p(x) \log \frac{p(\mathbf{x})}{p^{\mathrm{ind}}(\mathbf{x})} \tag{14}$$

and in this specialized form is known as the mutual information.[42] As well appreciated, the KLD is not symmetric in the distributions. The form used above is the natural one to use in order to compare with an independent pdf. The KLD is used extensively[28,42,43,64,70] to compare different probability distributions. It is of interest to also consider the configurational entropy difference between that based on the true $p(\mathbf{x})$ and assumed independent $p^{\mathrm{ind}}(\mathbf{x})$ distributions

$$\Delta H \equiv H_{\mathrm{dep}} - H_{\mathrm{ind}} = -[\sum_{x\in X} p(\mathbf{x}) \log p(\mathbf{x}) - \sum_{x\in X} p^{\mathrm{ind}}(\mathbf{x}) \log p^{\mathrm{ind}}(\mathbf{x})] \tag{15}$$

as a measure of the loss of entropy arising from the dependences among the random variables. However, as show in Appendix A, there is the identity,

$$\sum_{x\in X} [p(\mathbf{x}) - p^{\mathrm{ind}}(\mathbf{x})] \log p^{\mathrm{ind}}(\mathbf{x}) = 0 \tag{16}$$

Thus

$$D(p\,\|\,p^{\mathrm{ind}}) = -\Delta H \tag{17}$$

and therefore the physical information is contained in either formulation.

2.4. Computational Workflow. The protocol for generating the KLD entropy of the states proceeds as follows, based on having an atomistic trajectory available. (1) With some set of dihedrals of interest in mind, generate from the simulation data a trajectory for these dihedral angles. (2) For each dihedral, evaluate means and standard deviations using conventional dihedral angle definitions over the trajectory. By definition of a conformer, the width of a conformer's

DOI: 10.1021/jp5102412
J. Phys. Chem. B 2015, 119, 3621−3634

probability distribution is relatively small compared with the separation between the conformers' means. Therefore, the corresponding standard deviation will be large for multiconformer dihedrals relative to the standard deviation for the one conformer dihedrals, and this can be used as a quick screen for the multiconformation dihedrals that can contribute to the configurational entropy. For these large standard deviation dihedrals, examine the corresponding histograms to confirm the analysis. The histograms may indicate the k-means number of cluster centers to use but, as explicitly demonstrated below, it is worth exploring how many cluster centers to use. (3) For an assignment of the number of clusters for each dihedral, use a k-means program with the Fisher metric, as outlined in Section 2.2, to find, for each dihedral of interest, the cluster centroids, their populations, and the configurational assignments trajectory that, for each snapshot, provides the discrete conformer index for each dihedral. (4) This discrete, configurational assignments trajectory of the N $1DH_n = \{1, 2, ..., M_n\}$ dihedrals provides the data for a states trajectory that it is convenient to write as a trajectory of N dimensional strings, $\mathrm{NDH}(t) = [m_1(t), m_2(t), ..., m_N(t)]$ $[m_i(t) \in M_n]$. The fractional occupations for these states are obtained by matching the number of occurrences in the trajectory of these strings to those of the $N_S = \prod_{n=1}^{N} M_n$ possible state strings, $\mathrm{NDH} = (m_1, m_2, ..., m_N)$ $(m_i \in M_n)$. The independent state probabilities are obtained by multiplying together the fractional conformational occurrences for each dihedral. In the Supporting Information, a program is given that constructs the N_S possible state strings and does the matching of the $\mathrm{NDH}(t)$ trajectory strings to these possible NDH states, and constructs the state probabilities and the state trajectory. (5) The state and independent state probabilities in the discrete configurational assignment trajectory file are used in the KLD expression in eq 14 to generate the relative entropy.

3. RESULTS

3.1. k-Means Fisher Clustering of a Synthetic Gaussian Trajectory. Before examining the MD trajectory data, it is worth analyzing synthetic data to see how histograms in angular space and clustering using a circular statistics-based metric can provide different "views" of clusters. The issue arises from the circular statistics definition of an mean angle and distance to the mean angle, required for centroid-based and other clustering methods.[61] The number of clusters that would result if an ad hoc procedure based on the angular histograms were used versus circular statistics-based clustering are not the same in general.

Points whose values are distributed as normal distributions $N(\mu,\sigma)$ with mean, μ, and standard deviation, σ, were generated using the Box−Mueller algorithm with 100000 points, and three of them were combined with different μ, σ values to form representative point "trajectories" in angle space. Three cases: A, B, and C, were designed to explore the connection between angular histograms and k-means Fisher clustering.

The parameters for these three cases are given in Table 1, and the corresponding histograms in angle space are displayed in Figures 1−3. k-Means Fisher clustering identifies in each case three clusters whose centers and populations are listed in Table 1. Figures 1−3 also present parametric XY plots of the trajectory points relative to the cluster centroids:

Table 1. k-Means Fisher Clustering of Sums of Three Gaussian Distributions Using Three Clusters

case	μ, σ[a] (degrees)	cluster[b]	center (degrees)	population
A	−60, 15	1	−60.14	0.333
A	0, 15	0	0.151	0.333
A	60, 15	2	60.37	0.333
B	−60, 15	1	−59.92	0.333
B	60, 15	0	55.61	0.333
B	80, 15	2	84.58	0.333
C	−60, 15	1	−59.86	0.334
C	60, 10	0	61.19	0.481
C	90, 30	2	111.1	0.186

[a]Gaussian mean, μ, and standard deviation, σ, parameters. [b]The centroid cluster index order and μ order do not have to match.

$$X_i = \cos(\theta_i) - \mathrm{mX}[k] \quad (i = 1, 2, ...N_p)$$
$$Y_i = \sin(\theta_i) - \mathrm{mY}[k] \quad (k = 0, 1, 2) \tag{18}$$

with $\mathrm{mX}[k]$ and $\mathrm{mY}[k]$ the components of the converged centroids defined in eq 7. These plots provide a graphic demonstration that k-means can pull out the correct number of conformations, even though visual inspection of the dihedral pdf in angle space would most likely lead to an incorrect count of the number of conformations.

For case A, in Figure 1, with its relatively well-separated peaks, the cluster populations and centers mirror the angle histogram.

In case B, Figure 2, where "underlying" the angle histogram's positive angle peak with its composition of two overlapping Gaussians centered on 60° and 80° (Table 1), the clustering does resolve the overlap and all the populations are equal at 1/3. Presumably, based on a dihedral angle plot, case B would be classified as only two configurations. The XY plot, with its three distinct line segments, does clearly show the presence of three distinct states. Case C, Figure 3, was constructed to have a shoulder structure from the broader distribution centered on 90° as reflected in the angle histogram. Here, too, the clustering does resolve into three clusters. The left peak in the angle histogram has population 1/3 and its XY plot for cluster 1 is similar to that in Case B.

The shoulder peak population distribution reflects the "decision" that a clustering algorithm must make when the points are coming from overlapping distributions with differing width parameters; here, σ = 10° and 30°. Cases B and C do illustrate the importance of using a proper distance metric, as defined in Section 2.2, with which to do the k-means clustering. Fortunately, and typically, the extreme example of dihedral space overlap constructed in Case B should be the exception when using real data.

3.2. Dihedral Conformations for the Leucine Zipper Helical and Disordered N Terminus. The simulation data for the intact leucine zipper maintains the α-helical character of each monomer except for some fraying at the N and C terminal ends. The phi and psi angles for all residues excluding residues 1 and 2 at the N terminus all remain within the normal ranges for α-helical backbone dihedrals; (ϕ, ψ) around (−60 ± 15°, −45 ± 15°). In contrast, when the N terminal sequence (residues 1−8 of the N-terminus) is pushed out by MD restraints by the procedure given in Section 2.1, so that the residues toward the N terminus are separated by an additional ∼7 Å, the N terminal sequence

DOI: 10.1021/jp5102412
J. Phys. Chem. B 2015, 119, 3621−3634

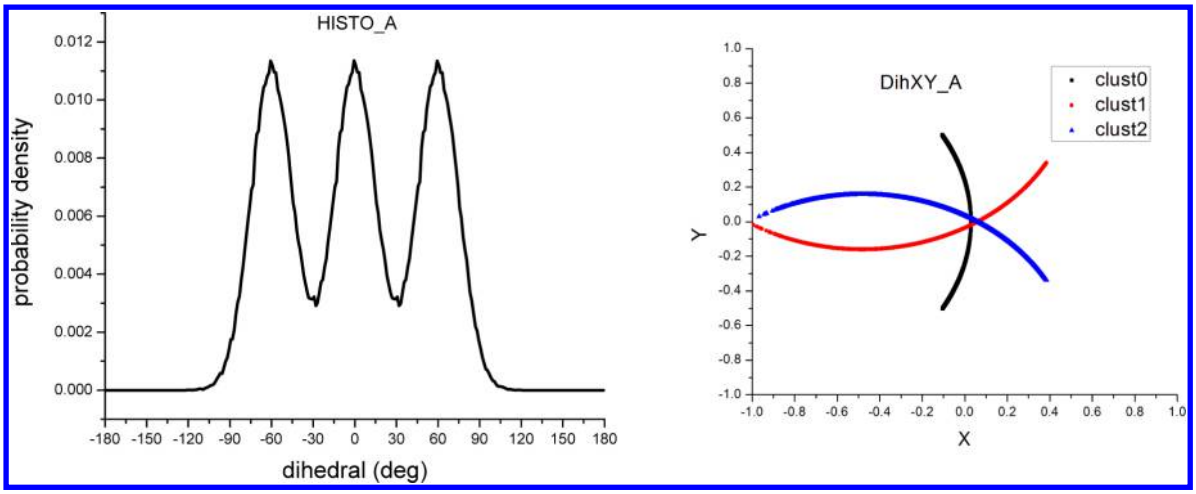


Figure 1. Case A: angle histogram generated from a sum of three normal distributions with means and widths given in Table 1 and XY parametric plot of the trajectory data (see eq 18 for the definitions of X and Y).

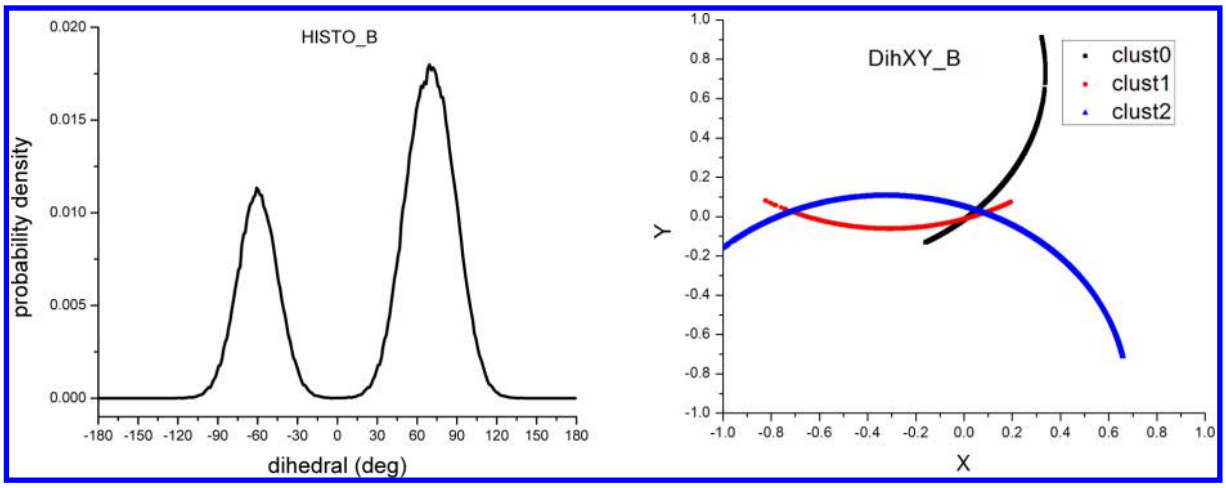


Figure 2. Case B: angle histogram generated from a sum of three normal distributions with means and widths given in Table 1 and XY parametric plot of trajectory data (see eq 18 for the definitions of X and Y).

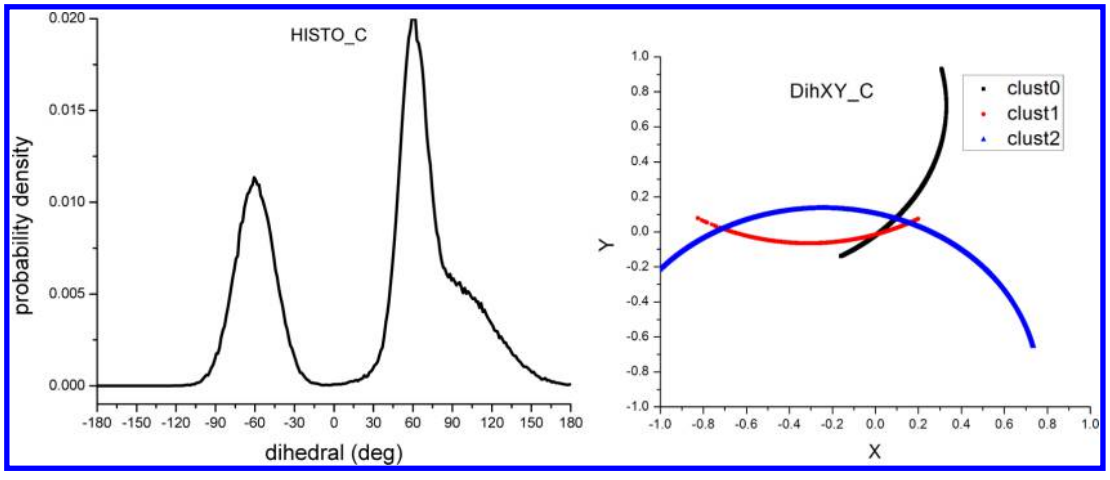


Figure 3. Case C. Angle histogram generated from a sum of three normal distributions with means and widths given in Table 1 and a XY parametric plot of trajectory data (see eq 18 for the definitions of X and Y).

partially loses its α-helical character. Figure 4 shows a snapshot where the residue range from LYS8 to the N terminus has lost its helix character.

These N terminal residues fluctuate among various conformations that are alpha helical about 50% of the time, based on the presence of the characteristic 1−4 hydrogen bonds between Met2-Glu6, Lys3-Asp7, and Gln4-Lys8.

DOI: 10.1021/jp5102412
J. Phys. Chem. B 2015, 119, 3621−3634

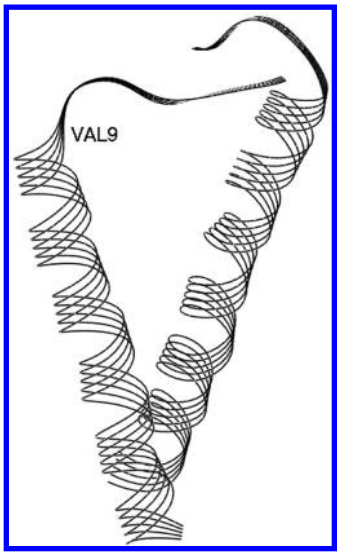


Figure 4. Separated dimer, accomplished with restraints, showing that the N terminal sequence melts (loses its α-helical character) to a certain extent (residues 3−8).

Such conformational fluctuations are best characterized via dihedral distributions. Figure 5 plots the mean and standard deviations of the N terminal sequence dihedrals and shows that residues 3−8 are most likely sampling more than one conformation, while the others sample only one conformation.

Figure 6 displays histograms of the residue 3, 4, and 5 phi and psi dihedrals to illustrate that these are sampling three conformers to form an IDP-like sequence in the N terminus.

The phi and psi dihedral histograms for residues 6−8 are similar in character, also exhibiting three conformer behavior, and confirm that Figure 5 points to the multi- versus one-conformation dihedrals.

3.3. States from Compositional Dihedral Clustering. Once the 1DH *k*-means Fisher dihedral clustering has been carried out to convergence, the dihedral angles can be assigned a discrete index, for each snapshot, providing a 1DH discretized trajectory for every dihedral. This trajectory is then used to obtain the state assignments in NDH according to the Cartesian product definition in eq 1. For the data discussed in Section 3.2, where the first 6 (residues 3−8) phi and psi dihedrals each have some population among three configurations, the state space has dimension 3^{12} = 531441. A plot of the populations of the first 100 states sorted by decreasing size is shown in Figure 7A, along with the assumed independent state populations for those states.

Figure 7B displays the cumulated state populations over the first 500 most populated states. From these plots it is evident that there is a strong dependence among the various states. It would be completely wrong to consider the dihedrals as independent random variables. Panel B shows that the first 100 (500) most populated states already account for ∼65% (∼90%) of the entire state population, again indicating the strong dependence among the states. These figures are based on 32 ns of trajectory data after equilibrating the separation of the monomers as described in Section 2.2. Another, independent 32 ns of trajectory data produced essentially identical results. Thus, the data displayed is characteristic of the N terminal sequence conformational fluctuations.

It is also of interest to explore the consequences of asserting that the 6 phi and psi dihedrals can be lumped into two, versus three, conformers, for the purposes of comparison with the above results and also for evaluation of the corresponding KLD entropy. Using the same data, but with a two-conformer-per-dihedral *k*-means assumption provides a total of 2^{12} = 4096 states. Figure 8 displays the sorted most probable state populations and the corresponding cumulating sum of populations. There is an analogous strong dependence among the dihedrals as for the three conformer per dihedral case. Here, the first 22 states out of the possible 4096 account for ∼90% of the total population. Again, the state population versus independent state population plots show the strong dependence among the dihedrals.

The results for both three and two conformers per dihedral consistently show that there is a strong dependence among the possible dihedral states. It should be noted that the ratio of the number of states $3^{12}/2^{12}$ ∼130 is an order of magnitude larger than that of the maximum probability ratio of the two to three conformers per dihedral, 0.25/0.035−7. On the basis of this reasoning, with three conformers per

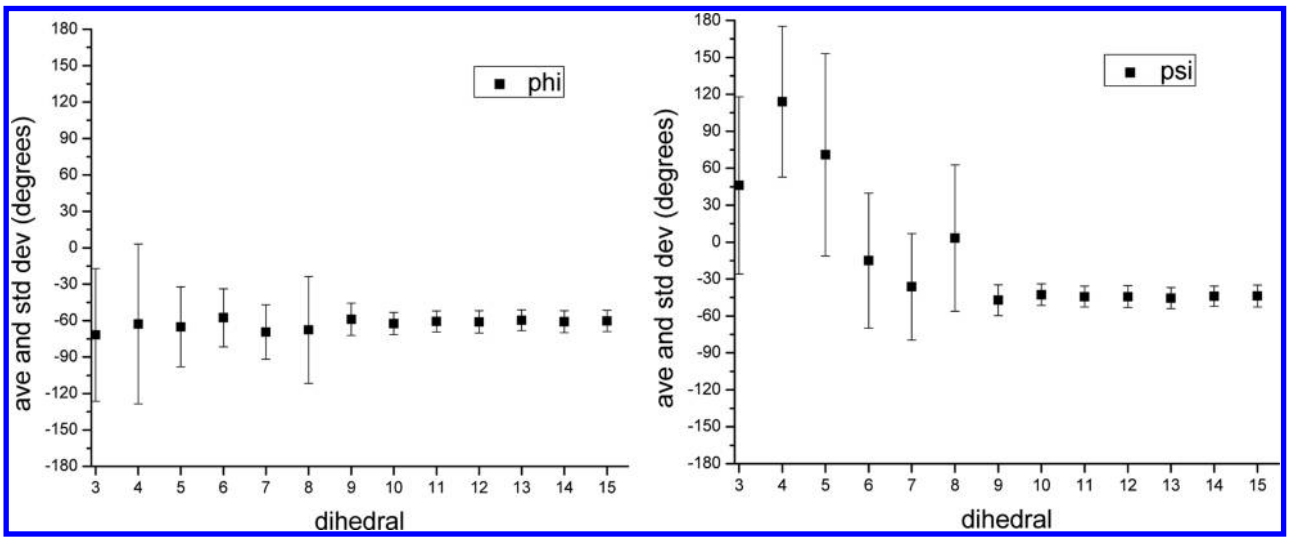


Figure 5. Mean and standard deviation of the phi and psi dihedrals of the monomer 1 N terminus (residues 3−15) of the leucine zipper. The residues 3−8 show more than one conformation behavior; the others remain in one conformation.

DOI: 10.1021/jp5102412
J. Phys. Chem. B 2015, 119, 3621−3634

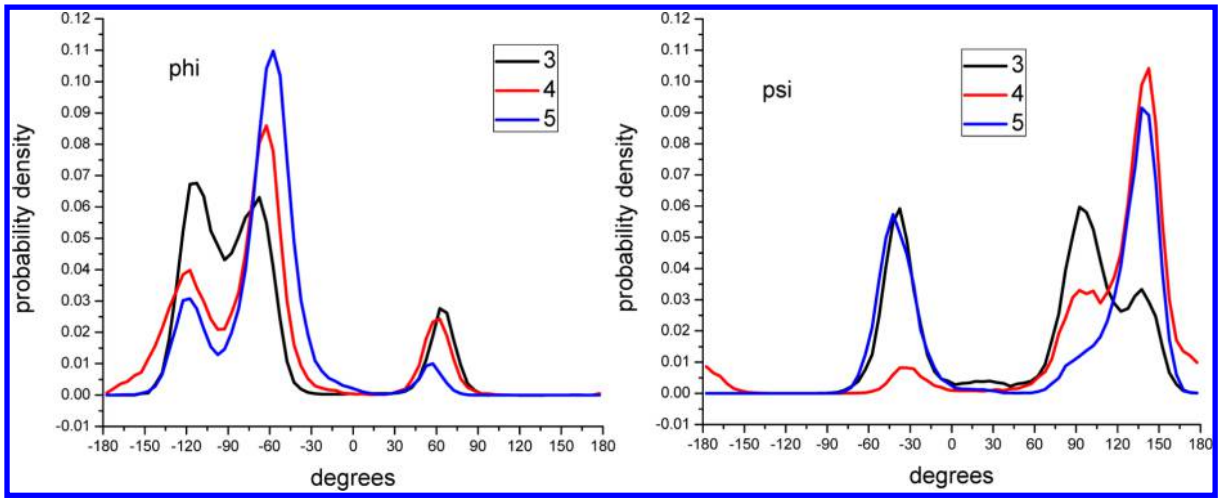


Figure 6. Histograms of the phi and psi backbone dihedrals of residues 3−5 that are fluctuating among three conformers for an IDP-like sequence of the N terminal part (residues 1−15) of the leucine zipper. Similar three peak histograms are obtained for residues 6−8, while the remaining 9−15 are all in one, α-helical conformer.

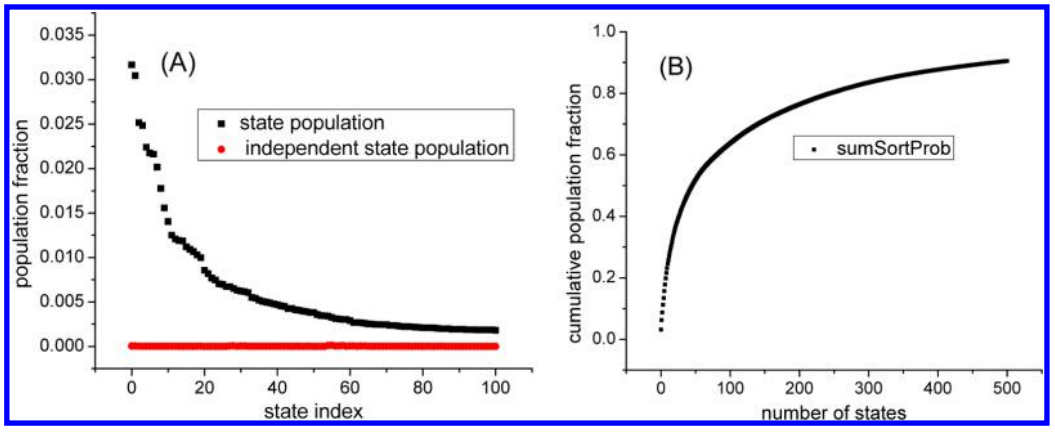


Figure 7. (A) Population fractions of the first 100 states sorted by decreasing size along with the assumed independent state populations for those states. There are 3^{12} = 531441 possible states for the 6 phi and 6 psi three-conformation dihedrals. The strong dependence among the dihedral conformers sampled is evident in this data representation. (B) Cumulated population fractions for the first 500 largest population states. The first 100 (500) most populated states account for ∼65% (∼90%) of the entire state population.

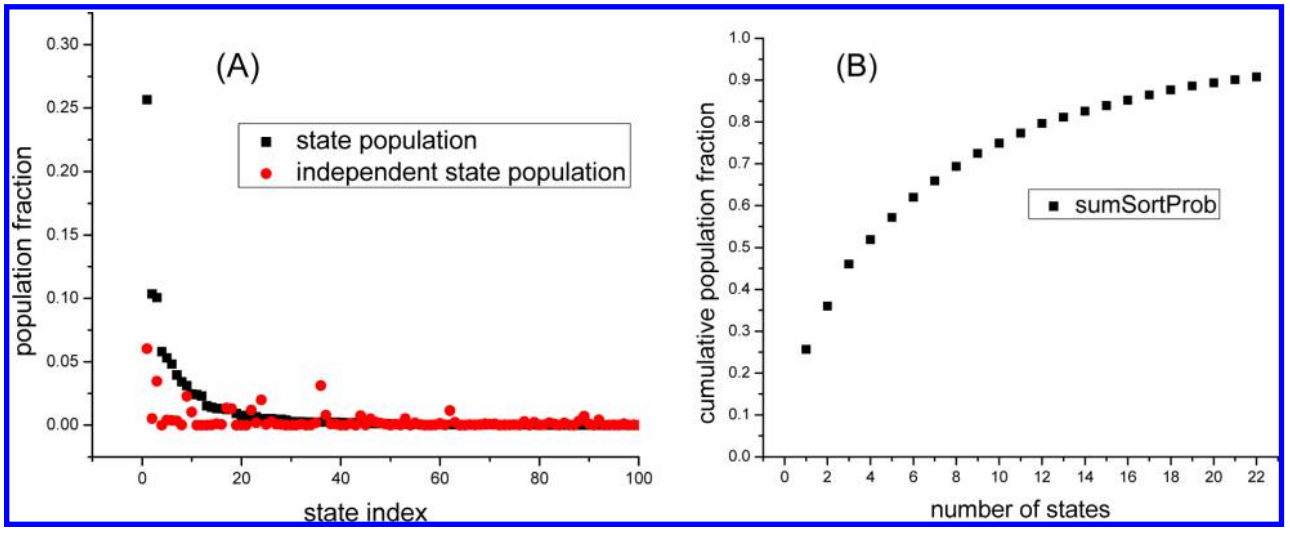


Figure 8. (A) Population fractions of the first 100 states sorted by decreasing size along with the assumed independent state populations for those states, out of the 2^{12} = 4096 possible states for the 6 phi and psi two-conformation dihedrals. The strong dependence among the dihedral conformers sampled is evident in this data representation. (B) Cumulated population fractions for the first 22 largest population states that account for ∼90% of the entire state population.

DOI: 10.1021/jp5102412
J. Phys. Chem. B 2015, 119, 3621−3634

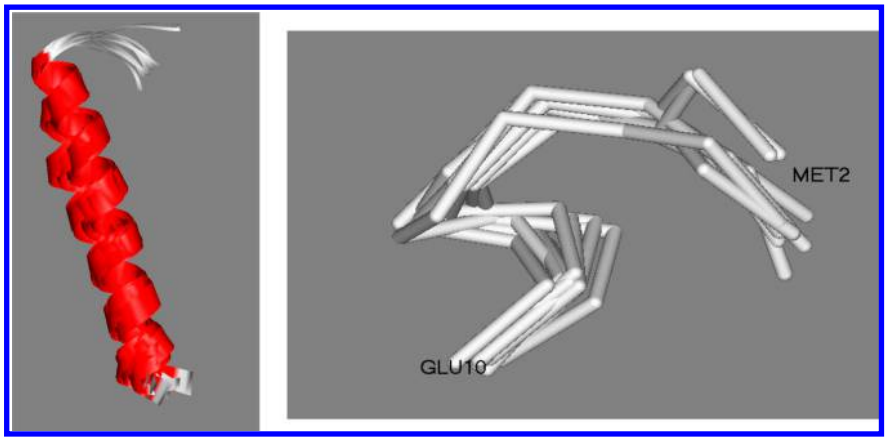


Figure 9. A selection of conformations of monomer 1 corresponding to the state with the highest probability of ∼0.03. The states are based on the conformations of residues 3−8 of monomer 1. Left: There is some diversity of conformers that is illusory as the figure includes residues 1 and 2 that are conformationally labile and not included in the state definition. Right: Conformations of residues 2−10. The nonhelical residues span Arg1 to Glu6.

dihedral, the extent of dependence is considerably larger than by assuming two conformers per dihedral.

While our emphasis is on the configurational entropy arising from multiconformational states, it is of interest to visualize the continuous space conformations of some of the states corresponding to the discrete states enumerated here. In particular, to contrast a high probability with a low probability state. Only monomer 1 is displayed because only its N terminal sequence became disordered in the simulation outlined in Section 2.1.

Figure 9 displays the representative monomer 1 conformations corresponding to the highest probability state, probability ∼ 0.03. (The states are based on conformations of residues 3−8 of monomer 1.) There is an apparent diversity of conformers because the figure includes residues 1 and 2 that are conformationally labile and are not included in the state definition. The right panel shows configurations excluding the first residue and indicates a reasonably consistent conformation for residues 3−8 that define the state. In all displayed snapshots, the nonhelical residues span Arg1 to Glu6. Examination of the trajectories for residues 3−8 for this state does confirm that the phi and psi dihedral angles are, for each angle, clustered around one conformer. It should be noted that the best fitting of snapshots, here carried out on all CA atoms, can itself lead to dispersion in the configurations and that small changes in dihedral space do cumulate.

Figure 10 displays all 10 snapshots from a low probability state, probability ∼0.0003. They are tightly clustered around a conformation where the non-α-helical residues span Arg1 to Val9, in contrast to the high probability state where resides 7−9 are helical. This comparison suggests that for this simulation data the loss in alpha helical character in the N terminus sequence does not propagate too far down the sequence.

3.4. Entropy of the N Terminus. As noted in Section 2.4 and proved in Appendix A, $D(p\|p^{\text{ind}}) = -\Delta H$ with $R\Delta H = \Delta S$ (R the gas constant), the difference in entropies between the true $p(\mathbf{x})$ and the assumed independent $p^{\text{ind}}(\mathbf{x})$ entropies defined in eq 15. Thus, ΔS measures the decrease of entropy that arises from the dependence among the dihedrals that are found in Section 3.3 relative to their assumed independence.

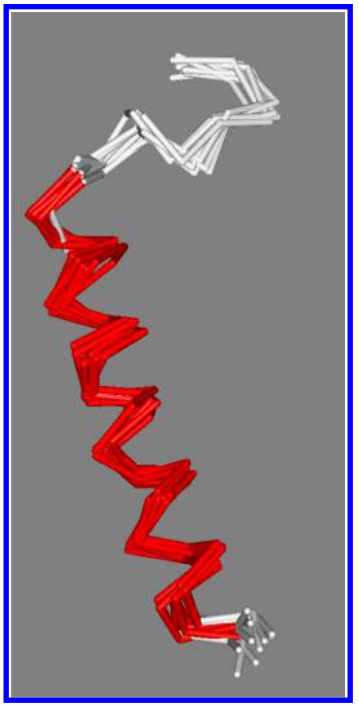


Figure 10. All conformations of monomer 1 corresponding to a low probability state (probability ∼0.0003). These conformations show loss of helical character from Val9 to Arg1. In contrast with the high probability state conformers displayed in Figure 9, here, the nonhelical residues span Arg1 to Val9.

Table 2 lists this decrease in entropy and its two contributions, expressed as a contribution to a free energy.

The A and B results in Table 2 were obtained from two independent trajectories, each with 32000 samples from 32 ns of data. As in Section 3.3, we assert that there are 12 dihedrals (6 phi and 6 psi) from residues 3−8 that are considered to have either three conformers or two conformers per dihedral. The histograms for residues 3−5 shown in Figure 6 and those for residues 6−8 (data not shown) are better described by three conformers for these dihedrals, though of course some conformer populations are quite low. The $T\Delta S$ values are consistent across the two data sets A and B. The TS_{ind} values are about double those from the true TS_{dep} values, indicating

DOI: 10.1021/jp5102412
J. Phys. Chem. B 2015, 119, 3621−3634

Table 2. Entropy Values for 12 Dihedrals with Three or Two Conformers Per Dihedral

data[a]	$T\Delta S$[b]	TS_{dep}[c]	TS_{ind}[d]
A. three per dihedral	−2.82	3.24	6.06
B. three per dihedral	−2.56	3.44	6.00
A. two per dihedral	−1.50	1.76	3.26
B. two per dihedral	−1.60	1.83	3.44

[a]On the basis of two trajectories, A and B, each with 32 ns of data and 32000 samples using k-means Fisher clustering with three or two centroids for each of the 12 dihedrals. [b]$T\Delta S = TS_{dep} - TS_{ind}$, TS_{dep} and TS_{ind} in kcal/mol at 300 K. [c]$TS_{dep} = -RT\sum_{\mathbf{x}\in\mathbf{X}} p(\mathbf{x}) \ln p(\mathbf{x})$ using the tate probabilities. [d]$TS_{ind} = -RT\sum_{\mathbf{x}\in\mathbf{X}} p^{ind}(\mathbf{x}) \ln p^{ind}(\mathbf{x})$ using the independent state probabilities.

the scale of the reduction in entropy relative to the assumed independent entropy.

Most importantly, the listed values of TS_{dep} are the free-energy contributions to entropy in excess of the zero value for the bound leucine zipper. They represent an entropic cost to transit from the disordered N terminal sequence to the dimerized leucine zipper. These contributions to the free energy are on the scale of binding free energies of proteins of 5−15 kcal/mol.

3.5. KLD Examples. In general, while the KLD does provide a quantitative measure of the difference between two probability distributions, $p(\mathbf{x})$ and $q(\mathbf{x})$, interpreting its magnitude is not straightforward. Here, because the reference is $q(\mathbf{x}) = p^{ind}(\mathbf{x})$ and then $D(p||p^{ind}) = -\Delta H$, the KLD provides the difference in the true and assumed independent entropies. As an attempt at providing some benchmark numbers that bear a resemblance to the results presented in Section 3.3 and Section 3.4, three examples of hypothetical state occupancies are developed and presented here. In general, of course, going from some nontrivial, many-dimensional probability distribution to its one-random-variable marginals required for the KLD measure based on independence is not feasible. For simplicity, we assume that each random variable takes on only two discrete values $X_i = \pm 1$, as for typical backbone dihedral conformations. Also assume that only $N + 1$ states have significant population, much like the data shown in Figure 8. Thus, out of the 2^N possible states for the two conformer per dihedral case, with probabilities $p_k(\mathbf{x})$ $(k = 1, 2, ..., 2^N)$, those with nonzero population can be enumerated as

$$p_1(\mathbf{x}) = (+1, +1, ..., +1)$$
$$p_2(\mathbf{x}) = (+1, +1, ..., -1)$$
$$.....$$
$$p_{N+1}(\mathbf{x}) = (-1, -1, ..., -1) \quad (19)$$

Then, as shown in Appendix B of the Supporting Information, the KLD values can be evaluated for the following three examples.

$$(1)\ p_k(\mathbf{x}) = \frac{1}{(N+1)} \quad (k = 1, 2, ..., N+1) \quad (20)$$

$$(2)\ p_k(\mathbf{x}) = \frac{2}{(N+1)(N+2)} \quad k(k = 1, 2, ..., N+1) \quad (21)$$

$$(3)\ p_k(\mathbf{x}) = \frac{1}{2}\left(\frac{1-r}{1-r^{N+1}}\right)[r^{k-1} + r^{N+1-k}]$$
$$(k = 1, 2, ..., N+1) \quad (22)$$

The first posits that all the nonzero probabilities are the same, the second that there is a linear variation, and the third that there is an exponential variation with "decay constant" of value $0 < r < 1$. The second and third examples do mirror the behavior of the state populations shown in Figures 7 and 8. The marginals and KLD expressions for example 1 are given in eqs (B.6), (B.7), and (B.12), for example 2 in (B.16), (B.17), and (B.18), and for example 3 in eqs (B.20)−(B.23), all of the Supporting Information. The values expressed as $-T\Delta S$ are plotted in Figure 11. All of them show a linear in

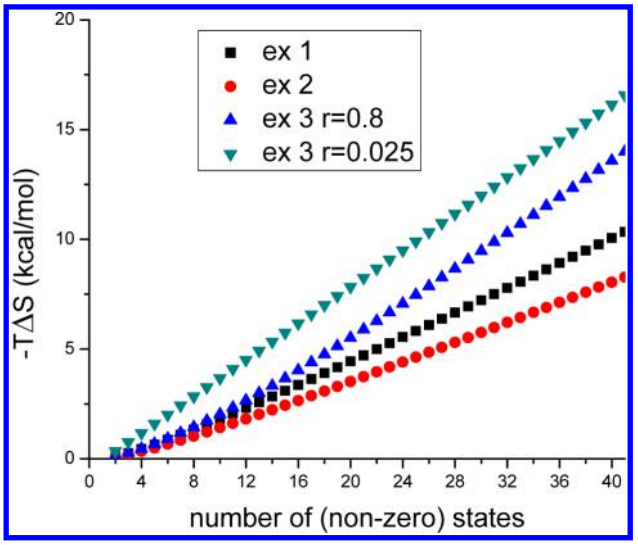


Figure 11. Values of $-T\Delta S$ for the three examples whose state probabilities p_k are given in eqs (3.3−3.5). Only $N + 1$ out of the possible 2^N states have nonzero probabilities. Example 3 depends on the parameter $0 < r < 1$ that measures the decay rate of the p_k's, having a faster dependence on r for small r.

N increase for sufficiently large N, as proved in the Supporting Information, Appendix B. For the two-configuration-per-dihedral data shown in Figure 8, there are on the order of 10 nonzero probability states, whose probabilities are falling rapidly. The $T\Delta S$ values are similar to those found in Figure 11 for all but example 3 with $r = 0.025$, a fast decay of the probability.

It is instructive to investigate the origin of the behavior in Figure 11 in example 3. The KLD numerator $\sum \mathbf{x} \in \mathbf{X} p(\mathbf{x}) \log p(\mathbf{x})$ "saturates" quickly with N and has a dependence on r. The p_k's have a greater dependence on the decay parameter r, at small r. For large r the dependence of p_k on k is slower, leading to more uniform p_k's; then, the corresponding entropy increases. But, this behavior gets swamped by the $\sum \mathbf{x} \in \mathbf{X} p(\mathbf{x}) \log p^{ind}(\mathbf{x})$ contribution that increases essentially linearly with N and is essentially independent of the r value.

4. DISCUSSION

In this work, a method to evaluate configurational entropy in a high dimensional space that directly focuses on the many dimensional probability distribution is introduced. This is in contrast with other methods[43,44,50,53] to incorporate dependencies that use cluster decomposition methods whereby, for

DOI: 10.1021/jp5102412
J. Phys. Chem. B 2015, 119, 3621−3634

example, the KLD for a set of dihedrals is expanded into contributions from successively larger numbers of dihedrals. While the chosen application is to dihedrals in a protein, the method is general. As has been long appreciated,[25−27] dihedrals are a natural coordinate set to use as their conformational space tends to be well-defined. Furthermore, "flips" of dihedral angles can lead to large overall conformational changes and then knowledge of the dependencies among the dihedrals and their state populations becomes particularly instructive. Our premise is that it is advantageous to cluster dihedrals compositionally rather than globally. That is, for dihedrals (or distances), each one is clustered to discrete space conformers, and states are defined from the Cartesian product of each dihedral's conformers.

Compositional, versus global, clustering should enhance the contrast among the so-defined states. In global clustering, a sum over the chosen coordinates is used to define states. For a given snapshot, there will be compromises among all the coordinates in the definition of the particular cluster that this snapshot falls in. With compositional clustering, each coordinate is assigned to a cluster, and the resulting composition should provide more sharply defined states. The distinction between the two clustering methods is simply rationalized by examining the limit case of, for example, N probability distribution functions of the form $p_n(x) = [\delta(x_n - X_+)+\delta(x_n - X_-)]/2$ with $X_\pm$ the peak locations. For N coordinates, each with this distribution, clustering will be the same when evaluated globally or compositionally. Either clustering method must pick out the same 2^N possible states. For dihedrals, if they really are very sharply restricted [i.e., for dihedral distributions consisting of two peaks that are each very narrow compared with their separation (the limit being the above distribution)] then the difference between the clustering using the two methods becomes smaller.

It is worth emphasizing that, for high-dimensional data, the global clustering becomes so computer intensive that only a sample of the MD data can be used as a practical matter. Even for our example with 12 dihedrals with three conformers per dihedral with its 531441 dimensional state space, using all the data (32000 snapshots) does not take significant computational resources.

The periodicity of dihedral angles does require definition of mean angles for the construction of the k-means centroids and their distances from the trajectory dihedral snapshots. That was resolved by introducing the Fisher definition of circular means. As discussed in Section 3.1 via a set of examples, what can be considered as ambiguous cluster numbers in angle space are well-resolved into clusters using the k-means Fisher procedure presented in Section 2.1.

With this procedure in hand, dihedral trajectories for the N terminus of a leucine zipper were analyzed. Our restrained simulation method did push one monomer away from the other in a way that the additional separation was largest toward the N terminus. Relative to the bound leucine zipper where all the N terminal sequence dihedrals, residues 3−15, sampled one conformer, residues 3−8 were each best-characterized by three conformers (with populations that varied widely from residue-to-residue). It is interesting that these backbone dihedrals sampled three versus two conformers. There is undoubtedly still influence from the other monomer on the conformers of the N terminal sequence. Examination of side chain interactions of the N terminus shows that there are monomer−monomer interactions, especially among ionized residues.

The k-means Fisher clustering provides more accurate resolution than angle space histograms for the MD data. For example, for the psi dihedral of residue 5 (Figure 6), there is one distinct conformer at a negative angle and what might be considered by "eye" to be one conformer at positive angle that, when subjected to k-means Fisher clustering, does resolve into two conformers. It should be noted that as long as the cluster assignment is accurate, low population conformers are properly accounted for in the construction of the state trajectory and state probabilities. The case C Gaussian example in Figure 3 does show unequivocally how shoulder-structured peaks in angle space will resolve into two clusters using this clustering method. These distinctions could be lost in global, versus compositional, clustering.

Because of the multiconfiguration character of the 6 phi and 6 psi dihedrals for residues 3−8, we defined the state space using these 12 dihedrals. The resulting state populations in Figure 7 show that while a variety of states are sampled, exemplifying IDP character, there are very strong propensities for particular states. The first 500 most populated states carry 90% of the states population, which corresponds to ∼0.1% of the possible states. While this is a very small percentage of the states, it is of course the case that relative to the most probable state (population ∼0.03), the 100th state is down in population by only 1 order of magnitude (population ∼0.003). Thus, many states are sampled, and the N terminal sequence can be considered an IDP. Similar considerations apply to the treatment of the data that assumes two conformers per dihedral.

The state probabilities will depend in a complicated fashion on the conformational probabilities of each dihedral. If independence is assumed then the state probabilities are given by all products of each dihedral's conformer probabilities. The more the conformer probabilities are weighted to one conformer at the expense of the others, the fewer high population states will result. Indeed, if all dihedrals were in one conformer, there would only be one state occupied. The data used was specifically focused on the dihedrals with robust multiconformer behavior as reflected in the small independent probabilities found.

Table 2 lists TS values in kcal/mol at T = 300 K to relate the entropic contributions to free energies. These values are relative to the zero TS contributions from the dimerized leucine zipper where all the dihedrals under consideration are found in one conformation, so there is only one state. Thus, the TS values in Table 2 are penalties required to form the stable dimer from the pushed-out N terminal sequence. The TS_{dep} values for the 12 dihedrals with three-configurations (two-configurations) per dihedral are around 3.3 (1.8) kcal/mol. It is of interest to compare these numbers to the largest possible entropies for 12 independent dihedrals with 3 (2) equally populated configurations of TS_{max} = 7.857 (4.957) kcal/mol at 300 K. Thus, in spite of the significant dependencies found, the entropy falls only by about a factor of 2.5. Clearly, these values can be as important as various energetic contributions to the overall free energy of stable zipper formation. Again, it is worth noting that the values of TS_{ind} relative to TS_{dep} are about a factor of 2 larger, and that these TS_{ind} values could be sufficiently large to be a dominant penalty for dimer formation. Thus, as an IDP, it may be useful to have some, but not too much, disorder to do the fine-

DOI: 10.1021/jp5102412
J. Phys. Chem. B 2015, 119, 3621−3634

tuning required for the competition between dimerized and dissociated states, particularly in response to various signals required for the timing of events such as translocation.

5. CONCLUDING REMARKS

The dihedral configurational entropy difference for the 12 dihedrals between the separated and bound N terminal sequence is a few kcal/mol, and this should be a typical magnitude for this number of degrees of freedom. It is therefore a potentially important contributor to the stability of IDPs interacting with their partners. The methods presented here can of course be applied to other IDPs that typically have loops that sample large configuration spaces until they interact with a partner.

The compositional clustering method can be used in contexts other than dihedrals. For example, distances between selected atoms can also be used to form an appropriate set of coordinates to carry out compositional clustering and may provide a more accurate definition of states. There are numerous proteins that undergo large conformational changes upon ligand binding. The role of entropy in going from a larger unbound configurational space to a more limited configurational space upon binding could be assessed by these methods.

The KLD measure that uses the total probability distribution provides a complete description of this entropic contribution to the free energy. Of course, the total probability distribution is in general not easy to obtain. The discretization scheme used here does reduce the state space considerably relative to a continuous space, though it still is an exponentially large space. There also is a contribution coming from the potential different widths of the underlying continuous space probability densities. The distinction between configurational and, in essence, vibrational entropy appropriate to hard degrees of freedom has been emphasized.[71] Entropic effects arising from differing widths of the probability densities for bond, bond angles, and dihedrals, when a ligand binds to a protein, have been assessed.[72] In the present context, if a particular IDP dihedral samples, for example, two resolved conformers, with each conformer characterized by a standard deviation σ_{IDP}, while the corresponding one-conformer dihedral probability density for the bound, structured case has a different standard deviation, σ_{BND}, then there will be a contribution to the entropy from this difference that is in addition to the configurational entropy. In particular, if $\sigma_{IDP} > \sigma_{BND}$, as should be the case for the less-constrained IDP versus bound, structured states, then the IDP entropy will be larger than the bound entropy. It would be of interest to assess and incorporate these contributions in the context of dependent dihedrals.

Finally, we note that the emphasis in this work has been on entropic contributions to binding free energies. From the atomistic trajectory, the time evolution of the states is also available. This reduced state space and time coarse-grained trajectory could be used to explore construction of Markov State Models[57,58] for protein folding and protein−protein interactions in general.

■ APPENDIX A. PROOF THAT $D(P||P^{IND}) = -\Delta H$

The Kullback Leibler divergence (KLD) between the two pdfs is

$$D(p\,\|\,p^{\mathrm{ind}}) = \sum_{x\in X} p(\mathbf{x})\log\frac{p(\mathbf{x})}{p^{\mathrm{ind}}(\mathbf{x})} \tag{A.1}$$

and the entropy difference is

$$\Delta H \equiv H_{\mathrm{dep}} - H_{\mathrm{ind}} = \sum_{x\in X} p(\mathbf{x})\log p(\mathbf{x}) - \sum_{x\in X} p^{\mathrm{ind}}(\mathbf{x})\log p^{\mathrm{ind}}(\mathbf{x}) \tag{A.2}$$

If

$$\sum_{x\in X}[p(\mathbf{x}) - p^{\mathrm{ind}}(\mathbf{x})]\log p^{\mathrm{ind}}(\mathbf{x}) = 0 \tag{A.3}$$

then

$$D(p\,\|\,p^{\mathrm{ind}}) = -\Delta H \tag{A.4}$$

The proof is notationally simplest with the assumption of each random variable taking on two possible values but is general. In particular, expressing $p(\mathbf{x})$ in terms of its 2^N possible states

$$\begin{aligned}
&\sum_{\mathbf{x}\in\mathbf{X}} p(\mathbf{x})\log p^{\mathrm{ind}}(\mathbf{x})\\
&= \sum_{\mathbf{x}\in\mathbf{X}} p(\mathbf{x})\log\prod_{n=1}^{N} p^{(n)}(x_n)\\
&= \sum_{s=1}^{2^N}\sum_{n=1}^{N} p_s\log p^{(n)}(x_n)\\
&= \sum_{n=1}^{N}\sum_{s=1}^{2^N} p_s\log p^{(n)}(x_n)\\
&= \sum_{n=1}^{N}\sum_{s=1}^{2^N} p_s\log p^{(n)}(X_n=+1) + \sum_{n=1}^{N}\sum_{s=1}^{2^N} p_s\log p^{(n)}(X_n=-1)
\end{aligned} \tag{A.5}$$

The definition of the nth singlet, marginal probability is

$$p^{(n)}(x_n) = \sum_{x\in\mathbf{X}\notin X_n} p(\mathbf{x}) \tag{A.6}$$

Using this definition shows that

$$\sum_{s=1}^{2^N} p_s(X_n=\pm 1) = \sum_{\mathbf{x}'\in\mathbf{X}',x_n=1} p_s \equiv p^n(X_n=\pm 1) \tag{A.7}$$

where the notation $\mathbf{x}'$ and $\mathbf{X}'$ excludes x_n and X_n in eq A.7. Therefore,

$$\begin{aligned}
&\sum_{\mathbf{x}\in\mathbf{X}} p(\mathbf{x})\log p^{\mathrm{ind}}(\mathbf{x})\\
&= \sum_{n=1}^{N}\sum_{s=1}^{2^N} p_s\log p^{(n)}(X_n=-1)\sum_{n=1}^{N}\sum_{s=1}^{2^N} p_s\log p^{(n)}(X_n=-1)\\
&= \sum_{n=1}^{N}[p^{(n)}(X_n=+1)p^{(n)}(X_n=+1) + p^{(n)}(X_n=-1)p^{(n)}(X_n=-1)]\\
&= \sum_{s=1}^{2^N} p_s^{\mathrm{ind}}\log p_s^{\mathrm{ind}}
\end{aligned} \tag{A.8}$$

and

DOI: 10.1021/jp5102412
J. Phys. Chem. B 2015, 119, 3621−3634

$$\sum_{x\in X} [p(\mathbf{x}) - p^{\mathrm{ind}}(\mathbf{x})]\log p^{\mathrm{ind}}(\mathbf{x}) = 0 \quad \text{(A.9)}$$

Consequently, from the definitions of $D(p\|p^{\mathrm{ind}})$ in eq A.1 and ΔH in eq A.2,

$$D(p\,\|\,p^{\mathrm{ind}}) = -\Delta \mathrm{H.\ Q.\ E.\ D} \quad \text{(A.10)}$$

■ ASSOCIATED CONTENT

Supporting Information

A program is provided that constructs the state probabilities and the state trajectory from a discretized dihedral trajectory. Derivations of the three KLD examples in Section 3.5 are given. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: cukier@chemistry.msu.edu. Tel: 517-355-9715, ext. 263. Fax: 517-353-1793.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Computational support from the Michigan State University High Performance Computing Center and from the Department of Chemistry is gratefully acknowledged.

■ REFERENCES

(1) Huyghues-Despointes, B. M. P.; Pace, C. N.; Englander, S. W.; Scholtz, J. M. Measuring the Conformational Stability of a Protein by Hydrogen Exchange. In *Methods in Molecular Biology; Protein Structure, Stability, and Folding*; Murphy, K. P., Ed.; Humana Press Inc: Totowa, NJ, Vol. *168*.

(2) Pace, C. N.; Shirley, B. A.; McNutt, M.; Gajiwala, K. Forces Contributing to the Conformational Stability of Proteins. *FASEB J.* **1996**, *10*, 75−83.

(3) Dill, K. A. Dominant Forces in Protein Folding. *Biochemistry* **1990**, *29*, 7133−7155.

(4) Dyson, H. J.; Wright, P. E. Intrinsically Unstructured Proteins and Their Functions. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197−208.

(5) Wright, P. E.; Dyson, H. J. Intrinsically Unstructured Proteins: Re-Assessing the Protein Structure-Function Paradigm. *J. Mol. Biol.* **1999**, *293*, 321−331.

(6) Wright, P. E.; Dyson, H. J. Linking Folding and Binding. *Curr. Opin. Struct. Biol.* **2009**, *19*, 31−38.

(7) Habchi, J.; Tompa, P.; Longhi, S.; Uversky, V. N. Introducing Protein Intrinsic Disorder. *Chem. Rev.* **2014**, *114*, 6561−6588.

(8) Uversky, V. N. Unusual Biophysics of Intrinsically Disordered Proteins. *Biochim. Biophys. Acta, Proteins Proteomics* **2013**, *1834*, 932−951.

(9) Stadler, A. M.; Stingaciu, L.; Radulescu, A.; Holderer, O.; Monkenbusch, M.; Biehl, R.; Richter, D. Internal Nanosecond Dynamics in the Intrinsically Disordered Myelin Basic Protein. *J. Am. Chem. Soc.* **2014**, *136*, 6987−6994.

(10) Drobnak, I.; De Jonge, N.; Haesaerts, S.; Vesnaver, G.; Loris, R.; Lah, J. Energetic Basis of Uncoupling Folding from Binding for an Intrinsically Disordered Protein. *J. Am. Chem. Soc.* **2013**, *135*, 1288−1294.

(11) Rogers, J. M.; Steward, A.; Clarke, J. Folding and Binding of an Intrinsically Disordered Protein: Fast, but Not 'Diffusion-Limited'. *J. Am. Chem. Soc.* **2013**, *135*, 1415−1422.

(12) Huang, Y. Q.; Liu, Z. R. Kinetic Advantage of Intrinsically Disordered Proteins in Coupled Folding-Binding Process: A Critical Assessment of the "Fly-Casting" Mechanism. *J. Mol. Biol.* **2009**, *393*, 1143−1159.

(13) Miller, M. The Importance of Being Flexible: The Case of Basic Region Leucine Zipper Transcriptional Regulators. *Curr. Protein Pept. Sci.* **2009**, *10*, 244−269.

(14) Missimer, J. H.; Dolenc, J.; Steinmetz, M. O.; van Gunsteren, W. F. Exploring the Trigger Sequence of the GCN4 Coiled-Coil: Biased Molecular Dynamics Resolves Apparent Inconsistencies in NMR Measurements. *Protein Sci.* **2010**, *19*, 2462−2474.

(15) Steinmetz, M. O.; Jelesarov, I.; Matousek, W. M.; Honnappa, S.; Jahnke, W.; Missimer, J. H.; Frank, S.; Alexandrescu, A. T.; Kammerer, R. A. Molecular Basis of Coiled-Coil Formation. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 7062−7067.

(16) Bunagan, M. R.; Cristian, L.; DeGrado, W. F.; Gai, F. Truncation of a Cross-Linked GCN4-P1 Coiled Coil Leads to Ultrafast Folding. *Biochemistry* **2006**, *45*, 10981−10986.

(17) Dragan, A. I.; Privalov, P. L. Unfolding of a Leucine Zipper is Not a Simple Two-State Transition. *J. Mol. Biol.* **2002**, *321*, 891−908.

(18) Matousek, W. M.; Ciani, B.; Fitch, C. A.; Garcia-Moreno, B.; Kammerer, R. A.; Alexandrescu, A. T. Electrostatic Contributions to the Stability of the Gcn4 Leucine Zipper Structure. *J. Mol. Biol.* **2007**, *374*, 206−219.

(19) Meisner, W. K.; Sosnick, T. R. Barrier-Limited, Microsecond Folding of a Stable Protein Measured with Hydrogen Exchange: Implications for Downhill Folding. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 15639−15644.

(20) Meisner, W. K.; Sosnick, T. R. Fast Folding of a Helical Protein Initiated by the Collision of Unstructured Chains. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 13478−13482.

(21) Nikolaev, Y.; Pervushin, K. NMR Spin State Exchange Spectroscopy Reveals Equilibrium of Two Distinct Conformations of Leucine Zipper GCN4 in Solution. *J. Am. Chem. Soc.* **2007**, *129*, 6461−6469.

(22) Burkhard, P.; Stetefeld, J.; Strelkov, S. V. Coiled Coils: A Highly Versatile Protein Folding Motif. *Trends Cell Biol.* **2001**, *11*, 82−88.

(23) Gruber, M.; Lupas, A. N. Historical Review: Another 50th Anniversary: New Periodicities in Coiled Coils. *Trends Biochem. Sci.* **2003**, *28*, 679−685.

(24) Moutevelis, E.; Woolfson, D. N. A Periodic Table of Coiled-Coil Protein Structures. *J. Mol. Biol.* **2009**, *385*, 726−732.

(25) Edholm, O.; Berendsen, H. J. C. Entropy Estimation from Simulations of Non-Diffusive Systems. *Mol. Phys.* **1984**, *51*, 1011−1028.

(26) Go, N.; Scheraga, H. A. Use of Classical Statistical-Mechanics in Treatment of Polymer-Chain Conformation. *Macromolecules* **1976**, *9*, 535−542.

(27) Go, N.; Scheraga, H. A. Analysis of the Contribution of Internal Vibrations to the Statistical Weights of Equilibrium Conformations of Macromolecules. *J. Chem. Phys.* **1969**, *51*, 4751−4767.

(28) Noe, F.; Horenko, I.; Schutte, C.; Smith, J. C. Hierarchical Analysis of Conformational Dynamics in Biomolecules: Transition Networks of Metastable States. *J. Chem. Phys.* **2007**, *126*, 155102.

(29) Cossio, P.; Laio, A.; Pietrucci, F. Which Similarity Measure Is Better for Analyzing Protein Structures in a Molecular Dynamics Trajectory? *Phys. Chem. Chem. Phys.* **2011**, *13*, 10421−10425.

(30) Zhou, T.; Caflisch, A. Distribution of Reciprocal of Interatomic Distances: A Fast Structural Metric. *J. Chem. Theory Comp.* **2012**, *8*, 2930−2937.

(31) Brandon, C.; Tooze, J. *Introduction to Protein Structure*, 2nd ed.; Garland Publishing: New York, 1999.

(32) Echeverria, I.; Makarov, D. E.; Papoian, G. A. Concerted Dihedral Rotations Give Rise to Internal Friction in Unfolded Proteins. *J. Am. Chem. Soc.* **2014**, *136*, 8708−8713.

(33) Fitzgerald, J. E.; Jha, A. K.; Sosnick, T. R.; Freed, K. F. Polypeptide Motions Are Dominated by Peptide Group Oscillations Resulting from Dihedral Angle Correlations between Nearest Neighbors†. *Biochemistry* **2006**, *46*, 669−682.

DOI: 10.1021/jp5102412
J. Phys. Chem. B 2015, 119, 3621−3634

(34) Helfand, E. Theory of the Kinetics of Conformational Transitions in Polymers. *J. Chem. Phys.* **1971**, *54*, 4651−4661.
(35) Li, D. W.; Bruschweiler, R. In Silico Relationship between Configurational Entropy and Soft Degrees of Freedom in Proteins and Peptides. *Phys. Rev. Lett.* **2009**, *102*.
(36) Higo, J.; Sugimoto, Y.; Wakabayashi, K.; Nakamura, H. Collective Motions of Myosin Head Derived from Backbone Molecular Dynamics and Combination with X-Ray Solution Scattering Data. *J. Comput. Chem.* **2001**, *22*, 1983−1994.
(37) Hnizdo, V.; Darian, E.; Fedorowicz, A.; Demchuk, E.; Li, S.; Singh, H. Nearest-Neighbor Nonparametric Method for Estimating the Configurational Entropy of Complex Molecules. *J. Comput. Chem.* **2007**, *28*, 655−668.
(38) Karplus, M.; Kushick, J. N. Method for Estimating the Configurational Entropy of Macromolecules. *Macromolecules* **1981**, *14*, 325−332.
(39) Numata, J.; Wan, M.; Knapp, E. W. Conformational Entropy of Biomolecules: Beyond the Quasi-Harmonic Approximation. In *Genome Informatics 2007*; Miyano, S.; DeLisi, C.; Holzhutter, H. G.; Kanehisa, M., Eds.; **2007**; Vol. *18*, pp 192−205.
(40) Schlitter, J. Estimation of Absolute and Relative Entropies of Macromolecules Using the Covariance-Matrix. *Chem. Phys. Lett.* **1993**, *215*, 617−621.
(41) McQuarrie, D. A. *Statistical Mechanics*; Harper & Row: New York, 1976.
(42) Cover, T. M.; Thomas, J. A. *Elements of Information Theory*; John Wiley & Sons: New York, 1991.
(43) McClendon, C. L.; Hua, L.; Barreiro, G.; Jacobson, M. P. Comparing Conformational Ensembles Using the Kullback−Leibler Divergence Expansion. *J. Chem. Theory Comp.* **2012**, *8*, 2115−2126.
(44) King, B. M.; Tidor, B. Mist: Maximum Information Spanning Trees for Dimension Reduction of Biological Data Sets. *Bioinformatics* **2009**, *25*, 1165−1172.
(45) King, B. M.; Silver, N. W.; Tidor, B. Efficient Calculation of Molecular Configurational Entropies Using an Information Theoretic Approximation. *J. Phys. Chem. B* **2012**, *116*, 2891−2904.
(46) Kasinath, V.; Sharp, K. A.; Wand, A. J. Microscopic Insights into the NMR Relaxation-Based Protein Conformational Entropy Meter. *J. Am. Chem. Soc.* **2013**, *135*, 15092−15100.
(47) Prabhu, N. V.; Lee, A. L.; Wand, A. J.; Sharp, K. A. Dynamics and Entropy of a Calmodulin-Peptide Complex Studied by NMR and Molecular Dynamics. *Biochemistry* **2003**, *42*, 562−570.
(48) Fenley, A. T.; Muddana, H. S.; Gilson, M. K. Entropy-Enthalpy Transduction Caused by Conformational Shifts Can Obscure the Forces Driving Protein-Ligand Binding. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 20006−20011.
(49) Fenley, A. T.; Killian, B. J.; Hnizdo, V.; Fedorowicz, A.; Sharp, D. S.; Gilson, M. K. Correlation as a Determinant of Configurational Entropy in Supramolecular and Protein Systems. *J. Phys. Chem. B* **2014**, *118*, 6447−6455.
(50) Killian, B. J.; Yundenfreund Kravitz, J.; Gilson, M. K. Extraction of Configurational Entropy from Molecular Simulations Via an Expansion Approximation. *J. Chem. Phys.* **2007**, *127*, 024107.
(51) McClendon, C. L.; Friedland, G.; Mobley, D. L.; Amirkhani, H.; Jacobson, M. P. Quantifying Correlations between Allosteric Sites in Thermodynamic Ensembles. *J. Chem. Theory Comp.* **2009**, *5*, 2486−2502.
(52) Killian, B. J.; Kravitz, J. Y.; Somani, S.; Dasgupta, P.; Pang, Y.-P.; Gilson, M. K. Configurational Entropy in Protein-Peptide Binding: Computational Study of Tsg101 Ubiquitin E2 Variant Domain with an Hiv-Derived Ptap Nonapeptide. *J. Mol. Biol.* **2009**, *389*, 315−335.
(53) Suarez, E.; Suarez, D. Multibody Local Approximation: Application to Conformational Entropy Calculations on Biomolecules. *J. Chem. Phys.* **2012**, *137*, 084115.
(54) Suarez, E.; Diaz, N.; Mendez, J.; Suarez, D. Cencalc: A Computational Tool for Conformational Entropy Calculations from Molecular Simulations. *J. Comput. Chem.* **2013**, *34*, 2041−2054.
(55) Suarez, D.; Diaz, N. Sampling Assessment for Molecular Simulations Using Conformational Entropy Calculations. *J. Chem. Theory Comput.* **2014**, *10*, 4718−4729.
(56) Kaufman, L.; Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*; Wiley-Interscience: New York, 1990.
(57) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. Msmbuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale. *J. Chem. Theory Comput.* **2011**, *7*, 3412−3419.
(58) Senne, M.; Trendelkamp-Schroer, B.; Mey, A. S. J. S.; Schütte, C.; Noé, F. Emma: A Software Package for Markov Model Building and Analysis. *J. Chem. Theory Comput.* **2012**, *8*, 2223−2238.
(59) Singh, G.; Tieleman, D. P. Atomistic Simulations of Wimley-White Pentapeptides: Sampling of Structure and Dynamics in Solution. *J. Chem. Theory Comput.* **2013**, *9*, 1657−1666.
(60) Shao, J. Y.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J. Chem. Theory Comput.* **2007**, *3*, 2312−2334.
(61) Fisher, N. I. *Statistical Analysis of Circular Data*; Cambridge University Press: Cambridge, England, 1993.
(62) Hamacher, K. Efficient Quantification of the Importance of Contacts for the Dynamical Stability of Proteins. *J. Comput. Chem.* **2011**, *32*, 810−815.
(63) McClendon, C. L.; Hua, L.; Barreiro, G.; Jacobson, M. P. Comparing Conformational Ensembles Using the Kullback-Leibler Divergence Expansion. *J. Chem. Theory Comput.* **2012**, *8*, 2115−2126.
(64) Wolfe, K. C.; Chirikjian, G. S. Quantitative Comparison of Conformational Ensembles. *Entropy* **2012**, *14*, 213−232.
(65) van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P. *Biomolecular Simulation: The Gromos96 Manual and User Guide*. Vdf hochschulverlag AG an der ETH: Zürich, 1996.
(66) Su, L.; Cukier, R. I. Hamiltonian Replica Exchange Method Studies of a Leucine Zipper Dimer. *J. Phys. Chem. B* **2009**, *113*, 9595−9605.
(67) O'Shea, E. K.; Klemm, J. D.; Kim, P. S.; Alber, T. X-Ray Structure of the GCN4 Leucine Zipper, a 2-Stranded, Parallel Coiled Coil. *Science* **1991**, *254*, 539−544.
(68) Lou, H.; Cukier, R. I. *Analyzer, 2.0*; Michigan State University: East Lansing, 2008.
(69) Kullback, S.; Leibler, R. A. On Information and Sufficiency. *Ann. Math Stat.* **1951**, *22*, 79−86.
(70) Hamprecht, F. A.; Peter, C.; Daura, X.; Thiel, W.; van Gunsteren, W. F. A Strategy for Analysis of (Molecular) Equilibrium Simulations: Configuration Space Density Estimation, Clustering, and Visualization. *J. Chem. Phys.* **2001**, *114*, 2079−2089.
(71) Karplus, M.; Ichiye, T.; Pettitt, B. M. Configurational Entropy of Native Proteins. *Biophys. J.* **1987**, *52*, 1083−1085.
(72) Chang, C.-e. A.; Chen, W.; Gilson, M. K. Ligand Configurational Entropy and Protein Binding. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1534−1539.

DOI: 10.1021/jp5102412
J. Phys. Chem. B 2015, 119, 3621−3634